1-1-2016

A comparison of spatial predictors when datasets could be very large

J R. Bradley
University of Missouri, Florida State University

Noel A. Cressie
University of Wollongong, ncressie@uow.edu.au

Tao Shi
Ohio State University

A comparison of spatial predictors when datasets could be very large

Abstract

In this article, we review and compare a number of methods of spatial prediction, where each method is viewed as an algorithm that processes spatial data. To demonstrate the breadth of available choices, we consider both traditional and more-recently-introduced spatial predictors. Specifically, in our exposition we review: traditional stationary kriging, smoothing splines, negative-exponential distance-weighting, fixed rank kriging, modified predictive processes, a stochastic partial differential equation approach, and lattice kriging. This comparison is meant to provide a service to practitioners wishing to decide between spatial predictors. Hence, we provide technical material for the unfamiliar, which includes the definition and motivation for each (deterministic and stochastic) spatial predictor. We use a benchmark dataset of CO2 data from NASA's AIRS instrument to address computational efficiencies that include CPU time and memory usage. Furthermore, the predictive performance of each spatial predictor is assessed empirically using a hold-out subset of the AIRS data.

Disciplines

Engineering | Science and Technology Studies

Publication Details

Bradley, J. R., Cressie, N. & Shi, T. (2016). A comparison of spatial predictors when datasets could be very large. Statistics Surveys, 10 100-131.

Statistics Surveys
Vol. 10 (2016) 100–131
ISSN: 1935-7516
DOI: 10.1214/16-SS115

A comparison of spatial predictors when datasets could be very large*

Jonathan R. Bradley†

Department of Statistics, University of Missouri, Columbia, MO
Department of Statistics, Florida State University, Tallahassee, FL
e-mail: bradley@stat.fsu.edu

Noel Cressie

National Institute for Applied Statistics Research Australia,
University of Wollongong, Australia

and

Tao Shi

Department of Statistics, The Ohio State University, Columbus, OH

Abstract: In this article, we review and compare a number of methods of spatial prediction, where each method is viewed as an algorithm that processes spatial data. To demonstrate the breadth of available choices, we consider both traditional and more-recently-introduced spatial predictors. Specifically, in our exposition we review: traditional stationary kriging, smoothing splines, negative-exponential distance-weighting, fixed rank kriging, modified predictive processes, a stochastic partial differential equation approach, and lattice kriging. This comparison is meant to provide a service to practitioners wishing to decide between spatial predictors. Hence, we provide technical material for the unfamiliar, which includes the definition and motivation for each (deterministic and stochastic) spatial predictor. We use a benchmark dataset of CO_2 data from NASA's AIRS instrument to address computational efficiencies that include CPU time and memory usage. Furthermore, the predictive performance of each spatial predictor is assessed empirically using a hold-out subset of the AIRS data.

MSC 2010 subject classifications: Primary 62H11; secondary 62P12.
Keywords and phrases: Best linear unbiased predictor, GIS, massive data, reduced rank statistical models, model selection.

Received October 2014.

*This is an original survey paper.

†The authors would like to thank three anonymous referees and the editor for their constructive comments. We would also like to thank Andrew Zammit Mangion for his helpful discussions. This research was partially supported by NASA's Earth Science Technology Office through its Advanced Information Systems Technology program, under Grant NNH11ZDA001N-AIST. It was also partially supported by the U.S. National Science Foundation (NSF) and the U.S. Census Bureau under NSF grant SES-1132031, funded through the NSF-Census Research Network (NCRN) program, and by NSF grants DMS-1007060 and DMS-1308458. Finally, Cressie's research was also partially supported by a 2015-2017 Australian Research Council Discovery Grant, number DP150104576.

Contents

1. Introduction

We are in an era of "big data," where the sizes of available datasets are becoming increasingly larger. For example, consider datasets on weather from the National Oceanic and Atmospheric Administration (NOAA), on earnings from the US Census Bureau's Longitudinal Employer-Household Dynamics program, and on public health from the Centers for Disease Control and Prevention (CDC). In the commercial sector, big data is now available using technology that allows companies to gather information on a myriad of purchases [23]. Pharmaceutical organizations amass large amounts of drug-testing data through combinatorial chemistry, medium-to-high-throughput screening (HTS), and other new technologies [6]. Many of these datasets can be very large in size; for example, the National Aeronautics and Space Administration (NASA) collects millions of atmospheric CO_2 measurements per month over the globe using the Atmospheric Infrared Sounder (AIRS) instrument on the Aqua satellite and, more recently, from the Orbiting Carbon Observatory-2 satellite.

As a result, big data is an important and growing topic in statistics. In the spatial-data setting, there are additional challenges. For example, AIRS CO_2 data have global extent, but they are spatially sparse, depending on the time period of observation. Additionally, they exhibit complex spatial dependencies that may be nonstationary. Thus, the complexity of "big spatial data" has motivated many to propose new statistical methodologies for spatial prediction (e.g.,

see 15, Ch. 4; 44, for reviews). For example, there are methods that use separable covariance functions, tapered covariance matrices, composite likelihoods, and low-dimensional latent Gaussian processes. These methodologies are all motivated by the fact that the Gaussian likelihood is difficult to compute when the dataset is large. Specifically, the Gaussian likelihood involves the computation of an inverse and a determinant of an $n \times n$ covariance matrix, a task that is on the order of n^3 computations, where n represents the size of the spatial dataset.

Despite the growing number of spatial predictors that are becoming available, there has been no comprehensive comparison between (and among) both traditional and modern spatial predictors. Such a comparison would be highly useful to the more general scientific community. In particular, the GIS community often uses spatial interpolation and smoothing (e.g., see 29) and would benefit from such a comparison. We take the approach of specifying a common task, where we train prediction rules that will be scored in a well defined, universal way. Here, we shall review the parameterization, the algorithm, and the motivation of seven spatial predictors, also considered by Bradley et al. [4] in the context of local spatial predictor selection. The predictors chosen are representative of their genres.

We consider three traditional spatial predictors, namely traditional stationary kriging, smoothing splines, and negative-exponential distance-weighting; and we consider four more-recently-introduced spatial predictors, namely fixed rank kriging, one based on modified predictive processes, one based on a stochastic partial differential equation, and lattice kriging. Of these seven predictors, smoothing splines and distance-weighting represent the deterministic approach, and the remaining represent the stochastic approach. We use a benchmark dataset of CO_2 data from NASA's AIRS instrument to empirically compare the predictive performances, computation times, and memory usage of these spatial predictors.

Kriging based on a stationary covariance function has become a method of spatial prediction covered in standard textbooks [e.g., 1, 8, 11, 15, 40] and has a rich history [see 10, and the references therein]. Since this method of spatial prediction has become a staple, we consider it in our study of AIRS CO_2 data and call the approach *traditional stationary kriging* (TSK). Another common approach is spatial interpolation using splines, which is obtained by minimizing a penalized-least-squares criterion (e.g., see 35; 46). Hence, we also consider *smoothing splines* (SSP) in our comparisons.

However, both TSK and SSP are not practical for large datasets; for example, they cannot be computed for the entire AIRS dataset for computational reasons. One simple ad hoc solution to this "big data" problem is to use a spatial predictor based on *negative-exponential distance-weighting* (EDW) [see 11, p. 371, for a discussion on these deterministic approaches]. Here, a datum's negative log weight is proportional to the Euclidean distance from the prediction location to the datum's location (see Section 3.3 for more details on EDW).

Although EDW is computationally efficient, we are predominantly interested in spatial predictors that are derived from statistical models and are appropriate for big data. For example, low-rank statistical models provide a computationally

efficient way to obtain the optimal kriging predictor and associated measures of prediction error. For this reason, low-rank statistical modeling for spatially referenced data is a popular method in the literature. In the spatial univariate setting, see Cressie and Johannesson [12], Shi and Cressie [42], Banerjee et al. [2], Cressie and Johannesson [13], and Kang and Cressie [24]. In the spatio-temporal setting, see Wikle and Cressie [48], Wikle et al. [49], Cressie et al. [14], Kang et al. [25], Katzfuss and Cressie [27, 28], and Bradley et al. [5]. Among the possible low-rank spatial predictors, we include two in our comparison: *fixed rank kriging* (FRK), and the *modified predictive process* (MPP) approach.

FRK seeks efficient calculation of the kriging predictor in the setting where n is very-large-to-massive. An advantage of FRK is that the inverse of the covariance matrix can be achieved efficiently using the Sherman-Morrison-Woodbury identity [e.g., 22] (see Section 3.4 for more details). The approach taken by MPP is similar and starts by first predicting a low-rank random effect called the predictive process. Then, predictions of a latent process are found by multiplying the prediction of the random effect by a set of basis functions (see Section 3.5 for more details). Some have criticized the use of a low-rank representation of a latent Gaussian process and believe that in many settings much of the variability occurs at high frequencies (see 30; 43, for discussions). However, it should be noted that high-frequency or discontinuous basis functions can address this criticism, making low-rank methods suitable for predicting processes with mixed variability (see the rejoinder of 4).

The remaining two spatial predictors we considered impose parametric assumptions on the precision matrix of the latent random process. One is based on a *stochastic partial differential equation* (SPD) approach proposed by Lindgren et al. [30], and the other is *lattice kriging* (LTK) proposed by Nychka et al. [36]. Computational efficiency is achieved by assuming sparseness of the spatial precision matrix of the random-effects vector (see Sections 3.6–3.7 for more details).

In Section 2, we provide a brief review of the spatial-predictor methodologies that are currently available, and we discuss the clear need to use benchmark datasets within this literature. That is, with the variety of spatial predictors available, benchmark datasets are essential in determining their relative usefulness, since one can compare both the predictive and computational performances of competing spatial predictors. In Section 3, we present the seven methods of spatial prediction, ranging from the classical to the more recent ones designed to handle very-large-to-massive datasets; both deterministic and stochastic spatial predictors are considered. Details surrounding the predictors are presented systematically, along with the motivation behind each spatial predictor. In Section 4, we apply and compare these predictors using different-sized datasets of remotely sensed mid-tropospheric CO_2 measurements from NASA's AIRS instrument. We include the computation time and memory usage of each predictor in the comparison, along with an empirical comparison of predictive performance using a hold-out dataset. A concluding discussion is provided in Section 5, followed by two short technical appendices.

2. Motivation: Benchmarking spatial predictor methodologies

The goal of this article is to review and compare some of the more recently introduced spatial predictors that are competitors for handling very large spatial datasets. Hence, in our exposition, many spatial predictors that were developed in an era of smaller datasets are not compared. Nevertheless, we think it is worthwhile here to give a brief review of the spatial predictors that were prominent in that era.

Our original source for this brief review is Cressie [11, Section 5.9], where stochastic and non-stochastic spatial predictors are compared. Of the seven predictors compared in this article and presented in Section 3, two (SSP and EDW) are deterministic (i.e., non-stochastic). Initially, one might think that comparisons of deterministic and stochastic predictors would be of "apples and oranges," but Bradley et al. [4] and this manuscript show that when viewed as algorithms that take data as input and produce spatial predictors as output, there are "figures of merit" that can be applied to all predictors.

In the last 25 years, spatial mapping has become ubiquitous in our daily lives, powered by map-querying software that gives us our present location and a path (via private vehicle, public transport, walking, etc.) to a chosen location. Remote sensing has filled in most of the holes in global mapping, however Earth is a dynamic environment, particularly for geophysical variables such as greenhouse gases. Satellites have well specified orbits that result in spatially incomplete data in any one day, one week, or one month for that matter. Paradoxically, the more data we collect, the more we realize that large fractions of the globe are missing due to the orbit geometry. On top of this, instrument error results in an imperfect measurement (i.e., "noise"). Thus, spatial-prediction technology is needed that accounts for the "missingness" and the "noise" in potentially very large datasets.

In the 1960s, G. Matheron and L.S. Gandin almost simultaneously developed a spatial-prediction technology that recognized spatial correlation as a way to solve the problem of missingness. However, they did not handle the noise appropriately; that came much later [9]. Of the seven predictors given in Section 3, five are stochastic and built around the methodology developed by Matheron [32] and Gandin [18], and some account for the noise. Their predecessors are summarized in Cressie [11] and include: simple kriging, ordinary kriging, universal kriging, optimal estimation, kriging with intrinsic random functions, Markov random-field prediction, disjunctive kriging, and Bayesian spatial prediction.

The deterministic spatial predictors go back to Gauss [19]; Cressie [10] gives the context in an article that traces the origins of kriging. The two deterministic predictors we consider are joined by predictors such as spatial moving average, inverse-distance-weighted average, Delauney triangulation, Natural-Neighbour interpolation, splines, and multi-quadric biharmonic interpolation [11, Section 5.9].

In this article, we have chosen traditional stationary kriging (TSK) as a placebo spatial-prediction method, and the other six spatial predictors were chosen for their representativeness and for taking a more modern approach to

handling very large spatial datasets. For researchers who have other spatial predictors to add to our list of seven, we have made training and validation datasets available in the Supplemental Material, making it easy for new comparisons to be made. Section 3 contains the definitions and genesis of all seven spatial predictors. The vast literature available on performing spatial prediction gives reason and motivation for providing up-to-date comparisons of recently introduced predictors in practice. That is, as spatial predictors are introduced, reviews and comparisons that reflect the concerns of the modern researcher are helpful in solidifying a new spatial predictor's role among the many choices that are available. Since many of the aforementioned spatial predictors are motivated by practical considerations, we consider both the predictive performance and the practical performance of each spatial predictor.

We acknowledge that the practical performance of a predictor extends beyond statistical considerations. For example, the performance of each "spatial predictor technology" depends on the efficiency of the code used to compute the spatial predictor, both the computer and the software used to run the code, and the way in which the spatial predictor is implemented. For example, if MPP say is computed twice using two different scripts, one written in Matlab and another in R, then one would expect differences in the practical performance of the two different computations of the MPP predictor. This is partially due to the fact that Matlab has built-in multithreading for multicore compute machines, while R currently does not (e.g., see `http://www.mathworks.com/discovery/matlab-vs-r.html`). Additionally, many of these methods are traditionally fitted using different numerical algorithms. For example, MPP is often fitted using Markov Chain Monte Carlo (MCMC) with Metroplis updates [2], while the SPD approach is traditionally fitted using integrated nested Laplacian approximations (INLA), which was developed as a faster Bayesian alternative to MCMC [38]. Furthermore, there is a Bayesian version of FRK [24], but no public-use software is available.

Thus, there are a number of competing effects that may be used to explain why one predictor is "better" than another. From this perspective, our practical comparison is of the "spatial predictor technologies" provided by the statistics community, which we define as the combination of the choice of computer, software, code, statistical model, spatial predictor, and the method of fitting used to obtain the spatial predictor. In particular, MPP, SPD, and TSK all have R scripts that are open source [16, 37, 39], both SSP and FRK have Matlab code that is available, and EDW can be easily written in either language (we chose Matlab; see Appendix A for the Matlab code). Critically, our comparison is based on publicly available code, which is more useful to practitioners uninterested in writing/optimizing/debugging their own programs.

3. Seven spatial predictors

In this section, we provide details on the spatial predictors considered. They are: traditional stationary kriging (TSK), smoothing splines (SSP), negative-exponential distance-weighting (EDW), fixed rank kriging (FRK), the modified

predictive process approach (MPP), the SPDE approach (SPD), and lattice kriging (LTK). Notice that the spatial predictors could be deterministic or stochastic, and we have chosen several that have been proposed recently to handle big spatial datasets. Details of the seven predictors are set out according to: the parameterization associated with each spatial predictor; the algorithm used to compute the spatial predictor; and the motivation behind the spatial predictor.

Several of the spatial predictors that we consider can be motivated by a spatial mixed effects (SME) model [e.g., 12, 13]:

$$\text{Data Model}: \quad Z(\mathbf{u}) = Y(\mathbf{u}) + \epsilon(\mathbf{u}) \tag{3.1}$$

$$\text{Process Model}: \quad Y(\mathbf{u}) = \mu(\mathbf{u}) + \nu(\mathbf{u}) + \xi(\mathbf{u}); \;\; \mathbf{u} \in D, \tag{3.2}$$

where $Z(\cdot)$ represents data or potential data; $\epsilon(\cdot)$ represents measurement error; $\mu(\cdot)$ is a deterministic mean function; $\nu(\cdot)$ models small-scale variation; $\xi(\cdot)$ is a term that captures (often non-smooth) micro-scale variation; and $D \equiv \{\mathbf{u}_j : j = 1, \ldots, N\} \subset \mathbb{R}^d$ is a generic finite set of prediction locations. All stochastic components, $\epsilon(\cdot)$, $\nu(\cdot)$, and $\xi(\cdot)$ are assumed mutually independent. A very flexible way to represent $\nu(\cdot)$ is through a basis-function expansion,

$$\nu(\mathbf{u}) = \mathbf{S}_r(\mathbf{u})'\boldsymbol{\eta}; \;\; \mathbf{u} \in D, \tag{3.3}$$

where $\mathbf{S}_r(\cdot)$ is an r-dimensional vector of spatial basis functions and $\boldsymbol{\eta}$ is an r-dimensional vector of random coefficients.

The spatial random process $Z(\cdot)$ is observed over a subset of the spatial domain of interest $D \subset \mathbb{R}^d$; that is, $Z(\cdot)$ is observed at locations in the set $D_O \equiv \{\mathbf{s}_i : i = 1, \ldots, n\} \subset D$. The latent process $Y(\cdot)$ is of principal interest, and one wishes to predict it from the data $\{Z(\mathbf{s}) : \mathbf{s} \in D_O\}$. It is assumed that $\epsilon(\cdot)$ is a white-noise Gaussian process with mean zero and known $\text{var}(\epsilon(\cdot)) = \sigma_\epsilon^2 V_\epsilon(\cdot)$, where $V_\epsilon(\cdot) > 0$ is a known function that captures heteroskedasticity. Note that often variance estimates are obtained from measuring-instrument calibration and quality assurance [e.g., see 33, among others], in which case σ_ϵ^2 can also be considered as known. Let $\mu(\cdot) \equiv \mathbf{x}(\cdot)'\boldsymbol{\beta}$, where $\mathbf{x}(\mathbf{s})$ is a p-dimensional vector of known spatial covariates defined on all $\mathbf{s} \in D$, and $\boldsymbol{\beta}$ is a p-dimensional vector of unknown regression coefficients.

The low-rank representation of $\nu(\cdot)$ requires further explanation. For $i = 1, \ldots, r$, the i-th element of $\mathbf{S}_r(\cdot)$ is given by the function, $S_{i,r} : D \to \mathbb{R}$; and the r-dimensional random vector $\boldsymbol{\eta}$ is specified as a Gaussian process with mean zero and $r \times r$ covariance matrix $\mathbf{K}$. Finally, the random process $\xi(\cdot)$ is assumed to be a Gaussian white-noise process with mean zero and variance $\sigma_\xi^2 V_\xi(\cdot) > 0$, where $V_\xi(\cdot)$ is a known positive function.

It will be seen below that the SME model motivates some of the stochastic predictors, although clearly not so for the deterministic predictors. Critically, it is not our intention in this article to fit a single stochastic model given by (3.1) and (3.2); rather, we look at each of the spatial predictors algorithmically, as it acts on the data $\{Z(\mathbf{s}_i) : i = 1, \ldots, n\}$. We also consider a "central" spatial predictor for each prediction method, recognizing that embellishments may be needed in a particular application. Our goal is to make the review and comparison as straightforward and transparent as possible.

3.1. Traditional Stationary Kriging (TSK)

Its parameterization: The statistical model, from which TSK is an optimal spatial predictor, can be defined hierarchically. The data model is given by (3.1) with $V_\epsilon(\cdot) \equiv 1$, and σ_ϵ^2 is known. The process model is given by,

$$Y(\mathbf{u}) = \mathbf{x}(\mathbf{u})'\boldsymbol{\beta} + \nu(\mathbf{u}) + \xi(\mathbf{u}); \;\; \mathbf{u} \in D, \tag{3.4}$$

where $\mathbf{x}(\mathbf{u})$ is a p-dimensional vector of known spatial covariates that describes the large-scale variation, $\nu(\mathbf{u})$ represents small-scale variation, and independently $\xi(\mathbf{u})$ represents fine-scale variation.

The spatial random process $\nu(\cdot)$ is specified to have mean zero and a second-order stationary covariance function,

$$\text{cov}(\nu(\mathbf{u}+\mathbf{h}), \nu(\mathbf{u})) \equiv C(\mathbf{h}); \;\; \mathbf{h} \in \mathbb{R}^d, \tag{3.5}$$

where the function $C(\cdot)$ is positive-definite (e.g., Cressie, 1993, p.68). Specifically, in Section 4, we use the exponential covariance function given by,

$$C(\mathbf{h}) = \sigma_0^2 \text{exp}\left(-\frac{||\mathbf{h}||}{\theta}\right); \;\; \mathbf{h} \in \mathbb{R}^d, \tag{3.6}$$

where $\theta > 0$ and $\sigma_0^2 > 0$. Additionally, in Section 4 it is assumed that $V_\xi(\cdot) \equiv 1$. We organize these unknown parameters into the set $\boldsymbol{\theta}^{\text{TSK}} \equiv \{\boldsymbol{\beta}, \theta, \sigma_0^2, \sigma_\xi^2\}$.

The algorithm: To compute TSK for a given $\boldsymbol{\theta}^{\text{TSK}}$, first construct the $n \times n$ covariance matrix,

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}^{\text{TSK}}) \equiv \left(\text{cov}\left(\nu(\mathbf{s}_i), \nu(\mathbf{s}_j)|\theta, \sigma_0^2\right) : i, j = 1, \ldots, n\right) + \sigma_\xi^2 \mathbf{V}_\xi + \sigma_\epsilon^2 \mathbf{I}_n, \tag{3.7}$$

where the $n \times n$ diagonal matrix $\mathbf{V}_\xi \equiv \text{diag}(V_\xi(\mathbf{s}_i) : i = 1, \ldots, n)$, and $\mathbf{I}_n$ is the $n \times n$ identity matrix. Also construct the n-dimensional vector,

$$\text{cov}(\mathbf{Z}, Y(\mathbf{u})|\boldsymbol{\theta}^{\text{TSK}}) = \text{cov}(\mathbf{Z}, \nu(\mathbf{u})|\boldsymbol{\theta}^{\text{TSK}}) + \sigma_\xi^2 V_\xi(\mathbf{u})(I(\mathbf{u} = \mathbf{s}_1), \ldots, I(\mathbf{u} = \mathbf{s}_n))', \tag{3.8}$$

where $I(\cdot)$ represents the indicator function. Then define

$$\hat{Y}(\mathbf{u}, \mathbf{Z}|\boldsymbol{\theta}^{\text{TSK}}) \equiv \mathbf{x}(\mathbf{u})'\boldsymbol{\beta} + \text{cov}(\mathbf{Z}, Y(\mathbf{u})|\boldsymbol{\theta}^{\text{TSK}})'\boldsymbol{\Sigma}(\boldsymbol{\theta}^{\text{TSK}})^{-1}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}), \tag{3.9}$$

where $\mathbf{X} \equiv (\mathbf{x}(\mathbf{s}_1), \ldots, \mathbf{x}(\mathbf{s}_n))'$.

Modifying (3.9) to be a function only of the data $\mathbf{Z}$, we substitute in the ordinary least squares (OLS) estimate for $\boldsymbol{\beta}$ and maximum likelihood (ML) estimates of the covariance parameters where the likelihood assumes mean zero, covariance (3.7), and is based on detrended data [e.g., 11, p. 239 and pp. 291-292]. The estimated parameters are denoted as $\hat{\boldsymbol{\theta}}^{\text{TSK}}$, and TSK is defined by the predictor,

$$\hat{Y}^{\text{TSK}}(\mathbf{u}, \mathbf{Z}) \equiv \hat{Y}(\mathbf{u}, \mathbf{Z}|\hat{\boldsymbol{\theta}}^{\text{TSK}}); \; \mathbf{u} \in D. \tag{3.10}$$

To compute $\hat{Y}^{\text{TSK}}$, we use the R-package "geoR" version 1.7-4 [37]. The computational complexity of TSK is $O(n^3)$, and it has memory size that is $O(n^2)$.

The motivation: The spatial predictor given by (3.9) minimizes the mean squared prediction error,

$$E\left((Y(\mathbf{u}) - \hat{Y}(\mathbf{u}, \mathbf{Z}))^2 | \boldsymbol{\theta}^{\mathrm{TSK}}\right),$$

among the class of linear predictors, $\hat{Y}(\mathbf{u}, \mathbf{Z}) = \ell + \mathbf{k}'\mathbf{Z}$ (e.g., Cressie, 1993, Section 3.4.5).

3.2. Smoothing Splines (SSP)

Its parameterization: In our implementation of smoothing splines, there is a single parameter that trades off smoothness with goodness-of-fit, which we denote as $\theta^{\mathrm{SSP}} > 0$.

The algorithm: The smoothing spline predictor, for a given θ^{SSP}, is

$$\hat{Y}(\mathbf{u}, \mathbf{Z}|\theta^{\mathrm{SSP}}) \equiv \mathbf{x}(\mathbf{u})'\hat{\boldsymbol{\beta}}^{\mathrm{SSP}} + \mathbf{W}(\mathbf{u})'(\mathbf{W} + \theta^{\mathrm{SSP}}\mathbf{I}_n)^{-1}(\mathbf{Z} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{SSP}}), \quad (3.11)$$

where $\mathbf{x}(\mathbf{u})$ is a p-dimensional vector of known spatial covariates, $\mathbf{X} \equiv (\mathbf{x}(\mathbf{s}_1), \ldots, \mathbf{x}(\mathbf{s}_n))'$ is an $n \times p$ matrix, and

$$\hat{\boldsymbol{\beta}}^{\mathrm{SSP}} \equiv (\mathbf{X}'(\mathbf{W} + \theta^{\mathrm{SSP}}\mathbf{I}_n)^{-1}\mathbf{X})^{-1}\mathbf{X}'(\mathbf{W} + \theta^{\mathrm{SSP}}\mathbf{I}_n)^{-1}\mathbf{Z}.$$

In our implementation, the (i, j)-th entry of $\mathbf{W}$, say W_{ij}, is obtained from a radial basis function as follows,

$$||\mathbf{s}_i - \mathbf{s}_j||^2 \log\left(||\mathbf{s}_i - \mathbf{s}_j||\right), \quad (3.12)$$

and the n-dimensional vector $\mathbf{W}(\mathbf{u})$ has i-th entry, $||\mathbf{u} - \mathbf{s}_i||^2 \log\left(||\mathbf{u} - \mathbf{s}_i||\right)$ [e.g., 46, p. 31].

The value of θ^{SSP} is chosen based on minimizing a leave-one-out cross-validation error [46, pp. 47–52]. Denote this minimized value as $\hat{\theta}^{\mathrm{SSP}}$, and hence SSP is defined by the predictor,

$$\hat{Y}^{\mathrm{SSP}}(\mathbf{u}, \mathbf{Z}) \equiv \hat{Y}(\mathbf{u}, \mathbf{Z}|\hat{\theta}^{\mathrm{SSP}});\ \mathbf{u} \in D, \quad (3.13)$$

which is a function only of the data $\mathbf{Z}$. To compute $\hat{Y}^{\mathrm{SSP}}$, we use the Matlab (Version 8.0) function "griddata." The computational complexity of SSP is $O(n^3)$, and it has memory size that is $O(n^2)$.

The motivation: The parameter θ^{SSP} is used to achieve a balance between goodness-of-fit and degree-of-smoothness of the spatial predictor [46]. In $\mathbb{R}^2$, the smoothing spline predictor is the function $f(\cdot)$ that minimizes the following penalized sum of squares (Wahba, 1990, p.31; Nychka, 2001),

$$\frac{1}{n}\sum_{i=1}^{n}(Z(\mathbf{s}_i) - f(\mathbf{s}_i))^2 + \theta^{\mathrm{SSP}} \int\int \left(\frac{\partial^2 f(\mathbf{u})}{\partial^2 u_1} + 2\frac{\partial^2 f(\mathbf{u})}{\partial u_1 \partial u_2} + \frac{\partial^2 f(\mathbf{u})}{\partial^2 u_2}\right) du_1 du_2, \quad (3.14)$$

for $\mathbf{u} = (u_1, u_2)'$. Its generalization to $\mathbb{R}^d$ for any positive integer d, is straightforward.

3.3. Negative-Exponential Distance-Weighting (EDW)

Its parameterization: There is a single parameter used for controlling the weights in negative-exponential distance-weighting, which we denote as $\theta^{\mathrm{EDW}}>0$.

The algorithm: The data are weighted based on their Euclidean distance from the prediction location $\mathbf{u}$. Let $d_i(\mathbf{u}) \equiv ||\mathbf{u}-\mathbf{s}_i||$ be the Euclidean distance between $\mathbf{u}$ and $\mathbf{s}_i$. The negative-exponential distance-weighting predictor, for a given θ^{EDW}, is

$$\hat{Y}(\mathbf{u},\mathbf{Z}|\theta^{\mathrm{EDW}}) \equiv \frac{\sum_{i=1}^{n}\exp\{-\theta^{\mathrm{EDW}}d_i(\mathbf{u})\}Z(\mathbf{s}_i)}{\sum_{i=1}^{n}\exp\{-\theta^{\mathrm{EDW}}d_i(\mathbf{u})\}};\ \ \mathbf{u}\in D. \tag{3.15}$$

The value of θ^{EDW} is often prespecified in advance. In this article, we estimate θ^{EDW} by minimizing

$$\sum_{i=1}^{n}\left(Z(\mathbf{s}_i)-\hat{Y}(\mathbf{u},\mathbf{Z}|\theta^{\mathrm{EDW}})\right)^2.$$

Denote this minimized value as $\hat{\theta}^{\mathrm{EDW}}$, and hence EDW is defined by the predictor,

$$\hat{Y}^{\mathrm{EDW}}(\mathbf{u},\mathbf{Z}) \equiv \hat{Y}(\mathbf{u},\mathbf{Z}|\hat{\theta}^{\mathrm{EDW}});\ \mathbf{u}\in D. \tag{3.16}$$

To compute $\hat{\theta}^{\mathrm{EDW}}$, we use the Matlab (Version 8.0) function "fminsearch." Additionally, to compute $\hat{Y}^{\mathrm{EDW}}$, we wrote a simple MATLAB script (provided in the Appendix). The computational complexity of EDW is $O(nN)$, and it has memory size that is $O(nN)$, where recall that N is the number of locations in the entire spatial domain D.

The motivation: A datum closer to the prediction location gets more weight than one further away. Also, for a prediction location that is the same as a datum location, there can be positive weights on the other data.

3.4. Fixed Rank Kriging (FRK)

Its parameterization: The statistical model, from which FRK is derived as an optimal spatial predictor, can be defined hierarchically. The data model is given by (3.1) with both $V_\epsilon(\cdot)$ and σ^2_ϵ known. The process model is,

$$Y(\mathbf{u}) = \mathbf{x}(\mathbf{u})'\boldsymbol{\beta} + \mathbf{S}_r(\mathbf{u})'\boldsymbol{\eta} + \xi(\mathbf{u});\ \ \mathbf{u}\in D, \tag{3.17}$$

where $\mathbf{x}(\mathbf{u})$ is a p-dimensional vector of known spatial covariates that describes the large-scale variation, $\mathbf{S}_r(\mathbf{u})'\boldsymbol{\eta}$ represents small-scale variation, and independently $\xi(\mathbf{u})$ represents fine-scale variation. The p-dimensional vector $\boldsymbol{\beta}$, the r-dimensional random vector $\boldsymbol{\eta}$, and the Gaussian white-noise process $\xi(\cdot)$ are all defined below (3.3). In Section 4, it is assumed that $V_\xi(\cdot)\equiv 1$ and $\mathbf{S}_r(\cdot)\equiv \mathbf{S}^{\mathrm{BI}}_r(\cdot)$, an r-dimensional vector function of bisquare basis functions [e.g., 13]. Further,

the value of r is specified to be much smaller than n; as will be discussed at the end of this section, specifying $r \ll n$ leads to computational advantages.

We organize the unknown parameters into the set $\boldsymbol{\theta}^{\mathrm{FRK}} \equiv \{\boldsymbol{\beta}, \mathbf{K}, \sigma_\xi^2\}$.

The algorithm: Define the $n \times n$ matrix $\mathbf{V}_\epsilon \equiv \mathrm{diag}(V_\epsilon(\mathbf{s}_1), \ldots, V_\epsilon(\mathbf{s}_n))$ and the $n \times r$ matrix $\mathbf{S}_r^{\mathrm{BI}} \equiv (\mathbf{S}_r^{\mathrm{BI}}(\mathbf{s}_1), \ldots, \mathbf{S}_r^{\mathrm{BI}}(\mathbf{s}_n))'$. To compute FRK, for a given $\boldsymbol{\theta}^{\mathrm{FRK}}$, first construct the $n \times n$ covariance matrix,

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}^{\mathrm{FRK}}) \equiv \mathrm{cov}(\mathbf{Z}|\boldsymbol{\theta}^{\mathrm{FRK}}, \mathbf{S}_r^{\mathrm{BI}}) = \mathbf{S}_r^{\mathrm{BI}}\mathbf{K}(\mathbf{S}_r^{\mathrm{BI}})' + \sigma_\xi^2\mathbf{V}_\xi + \sigma_\epsilon^2\mathbf{V}_\epsilon,$$

where the $n \times n$ diagonal matrix $\mathbf{V}_\xi \equiv \mathrm{diag}(V_\xi(\mathbf{s}_i) : i = 1, \ldots, n)$. Also construct the n-dimensional vector,

$$\begin{aligned}&\mathrm{cov}(\mathbf{Z}, Y(\mathbf{u})|\boldsymbol{\theta}^{\mathrm{FRK}}, \mathbf{S}_r^{\mathrm{BI}})\\ &= \mathbf{S}_r^{\mathrm{BI}}\mathbf{K}\,\mathbf{S}_r^{\mathrm{BI}}(\mathbf{u}) + \sigma_\xi^2 V_\xi(\mathbf{u})(I(\mathbf{u} = \boldsymbol{s}_1), \ldots, I(\mathbf{u} = \boldsymbol{s}_n))';\ \ \mathbf{u} \in D,\end{aligned} \tag{3.18}$$

where recall that $I(\cdot)$ represents the indicator function. Then define

$$\begin{aligned}&\hat{Y}(\mathbf{u}, \mathbf{Z}|\boldsymbol{\theta}^{\mathrm{FRK}})\\ &\equiv \mathbf{x}(\mathbf{u})'\boldsymbol{\beta} + \mathrm{cov}(\mathbf{Z}, Y(\mathbf{u})|\boldsymbol{\theta}^{\mathrm{FRK}}, \mathbf{S}_r^{\mathrm{BI}})'\boldsymbol{\Sigma}(\boldsymbol{\theta}^{\mathrm{FRK}})^{-1}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta});\ \ \mathbf{u} \in D,\end{aligned} \tag{3.19}$$

where $\mathbf{X} \equiv (\mathbf{x}(\mathbf{s}_1), \ldots, \mathbf{x}(\mathbf{s}_n))'$.

Modifying (3.19) to be a function only of the data $\mathbf{Z}$, we substitute in the OLS estimate for $\boldsymbol{\beta}$ and the Expectation Maximization (EM) estimates of the covariance parameters; here the likelihood from which the EM estimates are obtained assumes that the detrended data follow a Gaussian distribution with mean zero and covariance (3.18) [26]. For a review of the EM algorithm in this setting, see Bradley et al. [3]. The estimated parameters are denoted as $\hat{\boldsymbol{\theta}}^{\mathrm{FRK}}$. Then FRK is defined by the predictor,

$$\hat{Y}^{\mathrm{FRK}}(\mathbf{u}, \mathbf{Z}) \equiv \hat{Y}(\mathbf{u}, \mathbf{Z}|\hat{\boldsymbol{\theta}}^{\mathrm{FRK}});\ \ \mathbf{u} \in D. \tag{3.20}$$

To compute $\hat{Y}^{\mathrm{FRK}}$, we use Matlab code that is available on the website http://niasra.uow.edu.au/cei/webprojects/UOW175995.html#1.

The motivation: The spatial predictor given by (3.19) minimizes the mean squared prediction error,

$$E\left((Y(\mathbf{u}) - \hat{Y}(\mathbf{u}, \mathbf{Z}))^2|\boldsymbol{\theta}^{\mathrm{FRK}}\right),$$

among the class of linear predictors, $\hat{Y}(\mathbf{u}, \mathbf{Z}) = \ell + \mathbf{k}'\mathbf{Z}$ [13].

The primary motivation for FRK, as described in Cressie and Johannesson [12, 13], is that $\boldsymbol{\Sigma}(\boldsymbol{\theta}^{\mathrm{FRK}})^{-1}$ can be computed efficiently using the Sherman-Morrison-Woodbury formula [e.g., 22]:

$$\begin{aligned}\boldsymbol{\Sigma}(\boldsymbol{\theta}^{\mathrm{FRK}})^{-1} &= (\sigma_\xi^2\mathbf{V}_\xi + \sigma_\epsilon^2\mathbf{V}_\epsilon)^{-1} - (\sigma_\xi^2\mathbf{V}_\xi + \sigma_\epsilon^2\mathbf{V}_\epsilon)^{-1}\mathbf{S}_r^{\mathrm{BI}}\\ &\quad \times \{\mathbf{K}^{-1} + (\mathbf{S}_r^{\mathrm{BI}})'(\sigma_\xi^2\mathbf{V}_\xi + \sigma_\epsilon^2\mathbf{V}_\epsilon)^{-1}\mathbf{S}_r^{\mathrm{BI}}\}^{-1}(\mathbf{S}_r^{\mathrm{BI}})'(\sigma_\xi^2\mathbf{V}_\xi + \sigma_\epsilon^2\mathbf{V}_\epsilon)^{-1}.\end{aligned} \tag{3.21}$$

Equation (3.21) allows efficient computation of $\boldsymbol{\Sigma}(\boldsymbol{\theta}^{\mathrm{FRK}})^{-1}$ in (3.19), since (3.21) involves inverses of $r \times r$ matrices and a diagonal $n \times n$ matrix. Specifically, the computational complexity of the right-hand side of (3.21) is $O(nr^2)$, which is linear in n [13]. Additionally, FRK has memory size that is $O(nr)$.

3.5. Modified Predictive Process approach (MPP)

Its parameterization: The statistical model, from which MPP is derived as an optimal spatial predictor, can be defined hierarchically. The data model is given by (3.1) with $V_\epsilon(\cdot) \equiv 1$, and σ_ϵ^2 is unknown. The process model is

$$Y(\mathbf{u}) = \mathbf{x}(\mathbf{u})'\boldsymbol{\beta} + \mathbf{S}_r^{\mathrm{PP}}(\mathbf{u}; \kappa, \sigma_\nu^2)'\boldsymbol{\eta} + \xi(\mathbf{u}); \quad \mathbf{u} \in D, \tag{3.22}$$

where $\mathbf{x}(\mathbf{u})$ is a p-dimensional vector of known spatial covariates, $\mathbf{S}_r^{\mathrm{PP}}(\mathbf{u}; \kappa, \sigma_\nu^2)'\boldsymbol{\eta}$ represents small-scale variability, both κ and σ_ν^2 are unknown parameters, and independently $\xi(\mathbf{u})$ represents fine-scale variability. The p-dimensional vector $\boldsymbol{\beta}$, the r-dimensional random vector $\boldsymbol{\eta}$, and the Gaussian white-noise process $\xi(\cdot)$ are all defined below (3.3).

Let $\{\mathbf{u}_1^*, \ldots, \mathbf{u}_r^*\} \equiv D^* \subset D$ be a set of ($r \ll n$) knots over the spatial domain D. The r-dimensional random vector $\boldsymbol{\eta}$ is taken to be Gaussian with mean zero and covariance matrix $\mathbf{K}^*$, where $\mathbf{K}^* \equiv \left\{C(\mathbf{u}_i^*, \mathbf{u}_j^*)\right\}$. The term $\mathbf{S}_r^{\mathrm{PP}}(\cdot; \kappa, \sigma_\nu^2)$ is an r-dimensional vector function defined as,

$$\mathbf{S}_r^{\mathrm{PP}}(\mathbf{u}; \kappa, \sigma_\nu^2)' \equiv \mathbf{k}(\mathbf{u})' \left(\mathbf{K}^*\right)^{-1}, \tag{3.23}$$

where $\mathbf{k}(\mathbf{u}) \equiv (C(\mathbf{u}, \mathbf{u}_i^*) : i = 1, \ldots, r)'$ also depends on parameters κ and σ_ν^2. In Section 4, $C(\mathbf{u}, \mathbf{v})$ is the exponential covariance function, a function of $||\mathbf{u} - \mathbf{v}||$, with scaling parameter $\kappa > 0$ and variance σ_ν^2.

The original predictive-process approach, proposed by Banerjee et al. [2], did not include a fine-scale variation process $\xi(\cdot)$, and this led to a variance of the hidden process that was underestimated. Later Finley et al. [17] modified the approach by introducing $\xi(\cdot)$, resulting in (3.22). The spatial random process $\xi(\cdot)$ is modeled as a mean-zero independent Gaussian process such that $\mathrm{var}(\xi(\mathbf{u})) = \sigma_\xi^2 V_\xi(\mathbf{u}) = C(\mathbf{u}, \mathbf{u}) - \mathbf{k}(\mathbf{u})' \left(\mathbf{K}^*\right)^{-1} \mathbf{k}(\mathbf{u})$. This leads to

$$\begin{aligned}
&\mathrm{var}\left(\mathbf{S}_r^{\mathrm{PP}}(\mathbf{u}; \kappa, \sigma_\nu^2)'\boldsymbol{\eta} + \xi(\mathbf{u})\right) \\
&= \mathbf{k}(\mathbf{u})' \left(\mathbf{K}^*\right)^{-1} \mathbf{k}(\mathbf{u}) + C(\mathbf{u}, \mathbf{u}) - \mathbf{k}(\mathbf{u})' \left(\mathbf{K}^*\right)^{-1} \mathbf{k}(\mathbf{u}) = C(\mathbf{u}, \mathbf{u}) = \mathrm{var}\left(Y(\mathbf{u})\right),
\end{aligned}$$

as it should be. We organize the unknown parameters into the set $\boldsymbol{\theta}^{\mathrm{MPP}} \equiv \{\boldsymbol{\beta}, \kappa, \sigma_\nu^2, \sigma_\epsilon^2\}$.

The algorithm: Markov Chain Monte Carlo (MCMC) techniques are used for inference on parameters in this setting [2, 17]. The prior distributions are taken as $\sigma_\nu^2 \sim \mathrm{IG}(a_\eta, b_\eta)$, $\kappa \sim \mathrm{U}(a_\kappa, b_\kappa)$, $\sigma_\epsilon^2 \sim \mathrm{IG}(a_\epsilon, b_\epsilon)$, and $\boldsymbol{\beta}$ has a flat prior, where σ_η^2, κ, σ_ϵ^2, and $\boldsymbol{\beta}$ are assumed mutually independent, $\mathrm{IG}(a, b)$ represents an

inverted gamma distribution with parameters a and b, and $\mathrm{U}(a,b)$ represents a uniform distribution with parameters a and b. Choices for the hyperparameters depend on the application, but in Section 3 we use the suggestions from Finley and Banerjee [16], who also give details of the MCMC computations.

We note here an undesirable feature of the MPP, namely that it predicts the process $Z(\cdot)$. Recall that the data model is given by,

$$Z(\mathbf{u}) = Y(\mathbf{u}) + \epsilon(\mathbf{u}); \;\; \mathbf{u} \in D, \tag{3.24}$$

and hence MPP predicts the process with the measurement error included. Consequently, MPP predictions will be exactly equal to the training data at training data locations $\{\mathbf{s}_i\}$, which is an undesirable property when $\sigma_\epsilon^2 > 0$. Typically, scientific interest is in $Y(\cdot)$ not in $Z(\cdot)$, and the measurement error (i.e., "noise") component $\epsilon(\cdot)$ in (3.24) should be filtered out.

The MCMC generates samples $\{Z(\mathbf{u})_1, \ldots, Z(\mathbf{u})_L\}$ from the posterior distribution of $Z(\mathbf{u})$. Then MPP is defined by the predictor,

$$\hat{Y}^{\mathrm{MPP}}(\mathbf{u}, \mathbf{Z}) \equiv \frac{1}{L} \sum_{\ell=1}^{L} Z(\mathbf{u})_\ell; \;\; \mathbf{u} \in D. \tag{3.25}$$

To compute $\hat{Y}^{\mathrm{MPP}}$, we use the R-package "spBayes" [16]. The computational complexity of MPP is $O(nLr)$, and it has memory size that is $O(nr)$.

The motivation: The spatial predictor given by (3.25) minimizes the mean squared prediction error,

$$E(Z(\mathbf{u}) - \hat{Y}(\mathbf{u}, \mathbf{Z}))^2; \;\; \mathbf{u} \in D, \tag{3.26}$$

where here the expectation is taken over $\mathbf{Z}$, $Z(\mathbf{u})$, and $\boldsymbol{\theta}^{\mathrm{MPP}}$. As we noted above, instead of $Y(\mathbf{u})$, the scientifically-less-interesting quantity $Z(\mathbf{u})$ appears in (3.26). The primary motivation of this approach is that since $r \ll n$, the Sherman-Morrison-Woodbury identity can be used to compute the precision matrix efficiently, and thus it was proposed for use with large spatial datasets.

3.6. SPDE approach (SPD)

Its parameterization: The statistical model, from which SPD is derived as an optimal spatial predictor, can be defined hierarchically. The data model is given by (3.1) with $V_\epsilon(\cdot) \equiv 1$, and σ_ϵ^2 is unknown. The process model is given by,

$$Y(\mathbf{u}) = \mathbf{x}(\mathbf{u})'\boldsymbol{\beta} + \mathbf{S}_r^{\mathrm{PL}}(\mathbf{u})'\boldsymbol{\eta}; \;\; \mathbf{u} \in D, \tag{3.27}$$

where $\mathbf{x}(\mathbf{u})$ is a p-dimensional vector of known spatial covariates that describes the large-scale variation, $\mathbf{S}_r^{\mathrm{PL}}(\mathbf{u})'\boldsymbol{\eta}$ represents small-scale variability, and notice in (3.27) that the fine-scale variability term $\xi(\cdot) \equiv 0$. The p-dimensional vector $\boldsymbol{\beta}$ and the $r(> n)$-dimensional vector $\boldsymbol{\eta}$ are defined below (3.3). Here the term

$\mathbf{S}_r^{\mathrm{PL}}(\cdot)$ is an r-dimensional vector function whose elements are piecewise-linear basis functions; and in contrast to FRK and MPP, $r > n$.

On the Euclidean space, define a set of r knots $\{\mathbf{u}_1^*, \ldots, \mathbf{u}_r^*\} \equiv D^*$, which contains the n locations of D_O; that is, $r > n$. The r-dimensional random vector $\boldsymbol{\eta}$ is specified to be a mean-zero Gaussian Markov random field (GMRF) defined on D^*. The precision matrix associated with $\boldsymbol{\eta}$ (i.e., $\mathbf{K}^{-1} \equiv \mathrm{cov}\,(\boldsymbol{\eta})^{-1}$) is based on parameters κ and σ_ν^2. The functional form of this precision matrix, and hence the neighborhood structure of the elements in $\boldsymbol{\eta}$, is found by solving a stochastic partial differential equation, which we describe below. We organize the unknown parameters into the set $\boldsymbol{\theta}^{\mathrm{SPD}} \equiv \{\boldsymbol{\beta}, \mathbf{K}^{-1}, \sigma_\epsilon^2\}$.

The algorithm: Bayesian inference proceeds without using MCMC; it is based on Integrated nested Laplacian approximations (INLA) in this setting [30, 38]. The goal of INLA is to approximate the marginal distribution,

$$\pi(Y(\mathbf{s})|\mathbf{Z}) = \int \pi(Y(\mathbf{s})|\boldsymbol{\theta}^{\mathrm{SPD}}, \mathbf{Z})\pi(\boldsymbol{\theta}^{\mathrm{SPD}}|\mathbf{Z})d\boldsymbol{\theta}^{\mathrm{SPD}}; \;\; \mathbf{s} \in D. \tag{3.28}$$

The strategy of INLA is to make nested approximations of (3.28); in particular, both $\pi(Y(\mathbf{s})|\boldsymbol{\theta}^{\mathrm{SPD}}, \mathbf{Z})$ and $\pi(\boldsymbol{\theta}^{\mathrm{SPD}}|\mathbf{Z})$ are approximated, which we shall, in general, denote as $\bar{\pi}(Y(\mathbf{s})|\boldsymbol{\theta}^{\mathrm{SPD}}, \mathbf{Z})$ and $\bar{\pi}(\boldsymbol{\theta}^{\mathrm{SPD}}|\mathbf{Z})$, respectively. The term $\pi(\boldsymbol{\theta}^{\mathrm{SPD}}|\mathbf{Z})$ is approximated using a version of the Laplace approximation from Tierney and Kadane [45]. Rue et al. [38] provide several choices to approximate $\pi(Y(\mathbf{s})|\boldsymbol{\theta}^{\mathrm{SPD}}, \mathbf{Z})$, namely a Gaussian approximation, a full Laplace approximation, and a simplified Laplace approximation. Their results suggest that the simplified Laplace approximation of $\pi(Y(\mathbf{s})|\boldsymbol{\theta}^{\mathrm{SPD}}, \mathbf{Z})$ is the most efficient and leads to accurate estimates. Hence, this is used in SPD. Finally, a discrete sum is used to approximate the integral in (3.28). That is,

$$\bar{\pi}(Y(\mathbf{s})|\mathbf{Z}) = \sum_k \bar{\pi}(Y(\mathbf{s})|\boldsymbol{\theta}_k^{\mathrm{SPD}}, \mathbf{Z})\bar{\pi}(\boldsymbol{\theta}_k^{\mathrm{SPD}}|\mathbf{Z})\Delta_k; \;\; \mathbf{s} \in D, \tag{3.29}$$

where $\{\Delta_k\}$ are "area weights." The choice of $\{\boldsymbol{\theta}_k^{\mathrm{SPD}}\}$ and $\{\Delta_k\}$ are made within the R-INLA package.

For the parameter model, prior distributions are chosen for $\boldsymbol{\theta}^{\mathrm{SPD}}$. As a default in the R-INLA package, $\boldsymbol{\beta} \sim \mathrm{Gau}(\mathbf{0}, \tau_\beta^2\mathbf{I})$, and $\log\left(1/\sigma_\nu^2\right)$, $\log\left(\sqrt{8}/\kappa\right)$, and $\log\left(\sigma_\epsilon^2\right)$ are distributed as Log-Gamma. Further, $\boldsymbol{\beta}$, σ_ν^2, κ, and σ_ϵ^2 are assumed to be mutually independent. The values of hyperparameters of the prior distribution are chosen heuristically [39, personal communication]; we used the default settings of the R-INLA package.

Denote the posterior probability density function of $Y(\mathbf{u})$ as $\pi(Y(\mathbf{u})|\mathbf{Z})$, and the INLA-approximated version as $\bar{\pi}(Y(\mathbf{u})|\mathbf{Z})$ [e.g., 38, Section 3]. Rejection sampling is then used to generate L values $\{Y(\mathbf{u})_1, \ldots, Y(\mathbf{u})_L\}$ from $\bar{\pi}(Y(\mathbf{u})|\mathbf{Z})$. Finally, the SPD spatial predictor is,

$$\hat{Y}^{\mathrm{SPD}}(\mathbf{u}, \mathbf{Z}) \equiv \frac{1}{L}\sum_{\ell=1}^{L} Y(\mathbf{u})_\ell; \;\; \mathbf{u} \in D. \tag{3.30}$$

To compute $\hat{Y}^{\text{SPD}}$, we use the R-package "inla" [38, 39]. The computational complexity of SPD is approximately $O(n)$, and the memory size is not available.

The motivation: The spatial predictor given by (3.30) minimizes the (approximate) posterior mean squared prediction error,

$$\int (Y(\mathbf{u}) - \hat{Y}(\mathbf{u}, \mathbf{Z}))^2 \bar{\pi}(Y(\mathbf{u})|\mathbf{Z}) dY(\mathbf{u}). \tag{3.31}$$

Computational efficiency is obtained through a connection between GMRFs and Gaussian processes that have a Matérn covariance function [31],

$$\frac{\sigma_\nu^2}{\Gamma(\alpha)2^{\alpha-1}}(\kappa||\mathbf{h}||)^\alpha K_\alpha(\kappa||\mathbf{h}||);\ \mathbf{h} \in \mathbb{R}^d, \tag{3.32}$$

where $K_\alpha(\cdot)$ is the modified Bessel function of the second kind of order $\alpha > 0$. Here, $0 < \alpha < \infty$ is a smoothing parameter, $\kappa > 0$ is a scaling parameter, and σ_ν^2 is the variance parameter.

A random process $\nu(\cdot)$ in $\mathbb{R}^d$ with covariance function given by (3.32) with $\sigma_\nu^2 = \Gamma(\alpha)\Gamma(\alpha + d/2)^{-1}(4\pi)^{-d/2}\kappa^{-2\alpha}$ is a solution to the following stochastic partial differential equation [47]:

$$(\kappa^2 - \Delta)^{\zeta/2}\nu(\mathbf{u}) = W(\mathbf{u});\ \mathbf{u} \in \mathbb{R}^d, \tag{3.33}$$

where $W(\cdot)$ is a Gaussian white-noise process with mean zero and variance 1, and $\zeta \equiv \alpha + d/2$ is a positive integer, $\kappa > 0$, and $\sigma_\nu^2 > 0$. In (3.33), the Laplacian Δ is defined by,

$$\Delta \equiv \sum_{i=1}^{d} \frac{\partial^2}{\partial^2 u_i}. \tag{3.34}$$

Now write $\nu(\mathbf{u}) = \mathbf{S}_r^{\text{PL}}(\mathbf{u})'\boldsymbol{\eta}$, and specify $\boldsymbol{\eta}$ to be a GMRF with precision matrix, $\text{cov}\,(\boldsymbol{\eta})^{-1}$. This precision is formed by solving (3.33) [30, Section 2.3], but not for all values of the smoothness parameter α; the solution can only be found for $\zeta(= \alpha + d/2)$ a positive integer.

Lindgren et al. [30] extend this modeling approach to handle nonstationarity by letting some of the parameters depend on spatial coordinates; they find the precision matrix associated with the random vector $\boldsymbol{\eta}$ that solves the following stochastic partial differential equation,

$$(\kappa(\mathbf{u})^2 - \Delta)^{\zeta/2}\{\sigma_\nu(\mathbf{u})^2\mathbf{S}_r^{\text{PL}}(\mathbf{u})'\boldsymbol{\eta}\} = W(\mathbf{u});\ \mathbf{u} \in \mathbb{R}^d, \tag{3.35}$$

where $\zeta \equiv \alpha + d/2$ is a positive integer, $\kappa(\mathbf{u}) > 0$, and $\sigma_\nu(\mathbf{u})^2 > 0$. Lindgren et al. [30] propose the model,

$$\log\left(\sigma_\nu(\mathbf{u})^2\right) \equiv \sum_i \beta_i^{(1)} B_i^{(1)}(\mathbf{u}) \tag{3.36}$$

and

$$\log\left(\kappa(\mathbf{u})^2\right) \equiv \sum_i \beta_i^{(2)} B_i^{(2)}(\mathbf{u}), \tag{3.37}$$

where $\{B_i^{(1)}(\cdot)\}$ and $\{B_i^{(2)}(\cdot)\}$ represent two different finite sets of smooth basis functions.

Finally, in our case of $\mathbb{R}^2$, $\mathbf{K}^{-1}$ is specified as follows: $\alpha = 1$ and hence $\zeta = 2$, since $d = 2$; $\{B_i^{(1)}(\cdot)\}$ is a set of four spherical basis functions of order three; and $\{B_i^{(2)}(\cdot)\}$ is a set of seven spherical basis function of order six [e.g., 30].

3.7. Lattice Kriging (LTK)

Its parameterization: The statistical model defining lattice kriging can be defined hierarchically. The data model is given by (3.1) with $V_\epsilon(\cdot) \equiv 1$; and σ_ϵ^2 is assumed known. The process model is given by,

$$Y(\mathbf{u}) = \mathbf{x}(\mathbf{u})'\boldsymbol{\beta} + \mathbf{S}_r(\mathbf{u})'\boldsymbol{\eta}; \;\; \mathbf{u} \in D, \tag{3.38}$$

where $\mathbf{x}(\mathbf{u})$ is a p-dimensional vector of known spatial covariates, $\mathbf{S}_r(\mathbf{u})'\boldsymbol{\eta}$ represents small-scale variability, and notice in (3.38) that the fine-scale variability term $\xi(\cdot) \equiv 0$. The p-dimensional vector $\boldsymbol{\beta}$ and the $r(> n)$-dimensional vector $\boldsymbol{\eta}$ are defined below (3.3). Here the term $\mathbf{S}_r(\cdot)$ is an r-dimensional vector of spatial basis functions. In Section 4, $\mathbf{S}_r(\cdot) \equiv \mathbf{S}_r^{\mathrm{WL}}$, an r-dimensional vector of "smooth" multiresolutional Wendland basis functions; notice that $r > n$.

We first give the single-resolution version of the spatial model. From Nychka et al. [36], define a set of r nodes $\{\mathbf{u}_1^*, \ldots, \mathbf{u}_r^*\} \equiv D^*$ on a regular grid contained in D and ordered in some pre-specified manner. Given this ordering, for the i-th node denote the set of its four-nearest neighbors as N_i. Then define the r-dimensional random vector $\boldsymbol{\eta} \equiv \mathbf{B}^{-1}\mathbf{e}$, where $\mathbf{e}$ is an r-dimensional Gaussian random vector with mean zero and variance $\sigma_\eta^2\mathbf{I}_r$. Note that $\mathbf{B}\boldsymbol{\eta} = \mathbf{e}$, which is the form of a simultaneous autoregressive (SAR) model, and

$$B_{i,j} = \begin{cases} 4 + \kappa^2, & \text{for } j = i \\ -1, & \text{for } j \in N_i \\ 0, & \text{elsewhere.} \end{cases} \tag{3.39}$$

Thus, in this single-resolution version, just one parameter κ^2 is used to describe the spatial dependence. For the LTK multi-resolution model with K resolutions, $\mathbf{B}$ is block diagonal with K blocks, each taking the same form as (3.39), and there are potentially K parameters, $\kappa_1^2, \ldots, \kappa_K^2$, that describe the spatial dependence. A feature of LTK and the block-diagonal form of $\mathbf{B}$ is the following. Consider two random effects in $\boldsymbol{\eta}$ corresponding to two basis functions at the same location, but from different resolutions. Then the LTK model assumes that the two random effects are conditionally independent; from a graphical perspective, this means that two nodes are created at the common location, but there is no edge between them, which seems counter-intuitive. We organize the unknown parameters into the set $\boldsymbol{\theta}^{\mathrm{LTK}} \equiv \{\boldsymbol{\beta}, \sigma_\eta^2, \kappa\}$.

The algorithm: Define $\mathbf{S}_r^{\mathrm{WL}} \equiv (\mathbf{S}^{\mathrm{WL}}(\mathbf{s}_1), \ldots, \mathbf{S}^{\mathrm{WL}}(\mathbf{s}_n))'$. To compute LTK, for a given $\boldsymbol{\theta}^{\mathrm{LTK}}$, first construct the $n \times n$ covariance matrix,

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}^{\mathrm{LTK}}) \equiv \mathrm{cov}(\mathbf{Z}|\mathbf{K}, \mathbf{S}_r^{\mathrm{WL}}) = \mathbf{S}_r^{\mathrm{WL}}\mathbf{K}(\mathbf{S}_r^{\mathrm{WL}})' + \sigma_\epsilon^2\mathbf{I}_n, \tag{3.40}$$

where recall that $\mathbf{I}_n$ is the $n \times n$ identity matrix and $\mathbf{K} \equiv \text{cov}(\boldsymbol{\eta})$. Also construct the n-dimensional vector,

$$\text{cov}(\mathbf{Z}, Y(\mathbf{u})|\boldsymbol{\theta}^{\text{LTK}}, \mathbf{S}_r^{\text{WL}}) = \mathbf{S}_r^{\text{WL}} \mathbf{K}\, \mathbf{S}_r^{\text{WL}}(\mathbf{u}).$$

Then define,

$$\begin{aligned}&\hat{Y}(\mathbf{u}, \mathbf{Z}|\boldsymbol{\theta}^{\text{LTK}}) \\ &\equiv \mathbf{x}(\mathbf{u})'\boldsymbol{\beta} + \text{cov}(\mathbf{Z}, Y(\mathbf{u})|\boldsymbol{\theta}^{\text{LTK}}, \mathbf{S}_r^{\text{WL}})'\boldsymbol{\Sigma}(\boldsymbol{\theta}^{\text{LTK}})^{-1}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}); \;\; \mathbf{u} \in D, \end{aligned} \tag{3.41}$$

where $\mathbf{X} \equiv (\mathbf{x}(\mathbf{s}_1), \ldots, \mathbf{x}(\mathbf{s}_n))'$.

Modifying (3.41) to be a function only of the data $\mathbf{Z}$, we substitute in the maximum likelihood estimate of $\boldsymbol{\theta}^{\text{LTK}}$ (denoted $\hat{\boldsymbol{\theta}}^{\text{LTK}}$). Then LTK is defined by the predictor,

$$\hat{Y}^{\text{LTK}}(\mathbf{u}, \mathbf{Z}) \equiv \hat{Y}(\mathbf{u}, \mathbf{Z}|\hat{\boldsymbol{\theta}}^{\text{LTK}}); \;\; \mathbf{u} \in D. \tag{3.42}$$

To compute $\hat{Y}^{\text{LTK}}$, we use the R package "LatticeKrig" [36]. The computational complexity and memory size are not known for LTK; we refer the reader to Figure 1 of Nychka et al. [36] for approximations.

The motivation: The spatial predictor given by (3.41) minimizes the mean squared prediction error,

$$E\left((Y(\mathbf{u}) - \hat{Y}(\mathbf{u}, \mathbf{Z}))^2|\boldsymbol{\theta}^{\text{LTK}}\right), \tag{3.43}$$

among the class of linear predictors, $\hat{Y}(\mathbf{u}, \mathbf{Z}) = \ell + \mathbf{k}'\mathbf{Z}$. A numerical motivation for LTK is that $\boldsymbol{\Sigma}(\boldsymbol{\theta}^{\text{LTK}})^{-1}$ can be found using sparse-matrix techniques [36].

4. A comparison of the seven spatial predictors: Mid-tropospheric CO_2 measurements

The Aqua satellite is part of the Earth Observing System (EOS), which is administered by the National Aeronautics and Space Administration (NASA). The Atmospheric Infrared Sounder (AIRS) is an instrument on board the Aqua satellite that retrieves information on atmospheric variables, amongst other things, CO_2. Specifically, the AIRS instrument collects CO_2 measurements in the form of spectra that are then converted to mid-tropospheric CO_2 values in parts per million (ppm) [7]. This type of global information has been used to great effect in raising public awareness on greenhouse gases and in determining policy regarding climate change (e.g., see https://www.ipcc.ch/).

The AIRS instrument records data over swaths (or paths) of Earth's surface (roughly 800 km wide) and extends from $-60°$ to $90°$ latitude. We use AIRS data retrieved from February 1 through February 9, 2010, which are converted into level-2 CO_2 data reported at a 17.6 km by 17.6 km spatial resolution.

The resulting AIRS CO_2 dataset consists of 74,361 total observations, which we shall use in a comparison of both the predictive performance and the computational performance of the seven spatial predictors. However, not every predictor can be computed using all 74,361 observations. For example, it is well

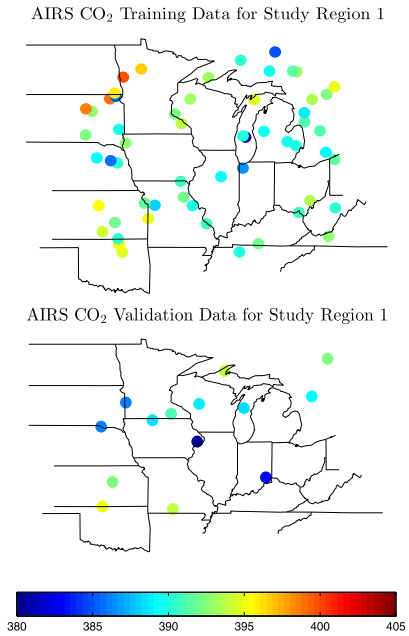

FIG 1. *A spatial dataset made up of 9 days of measurements of mid-tropospheric* CO_2 *in parts per million (ppm). The data considered are in North America between* $-49°$ *degrees and* $36°$ *degrees latitude and* $-80°$ *degrees and* $-99.5°$ *degrees longitude, from Februrary 1 through Februrary 9, 2010. The data are randomly split into training and validation datasets with* $n = 57$ *and* $m = 14$*, respectively.*

known that the traditional predictors, TSK and SSP, cannot handle datasets this large (or larger). Hence, we take subsets of the globe (i.e., D) that contain smaller numbers of data points than found in $\{Z(\mathbf{s}) : \mathbf{s} \in D_O\}$.

In Figure 1, we display Study Region 1, which covers the Midwest region of North America. Here, there is a total of just 71 observations, which we separate into two subsets of size $n = 57$ and $m = 14$. The n observations are the "training" data (top panel of Figure 1) used to fit the spatial predictors, and the m observations are the "validation" data used to assess the predictive performance of each spatial predictor (bottom panel of Figure 1); notice that we reserve roughly 20% of the data for validation. Our main reason for analyzing this small study region is to compare the predictive performance of *all* seven spatial predictors, which we do in Section 4.1. For this particular dataset, we found that trying other regions of the globe led to instability of one or more predictors. However, it is important to note that TSK and SSP can certainly be computed for a spatial datasets whose size is on the order of a thousand or less.

Although we are interested in comparing all the predictors, a number of them are designed to handle larger datasets. In particular, EDW, FRK, MPP, SPD, and LTK are relatively straightforward (but non-trivial) predictors that are intended for larger spatial datasets. Consider Study Region 2 in Figure 2, which covers the Americas and western Sahara between longitudes $-125°$ to $3°$

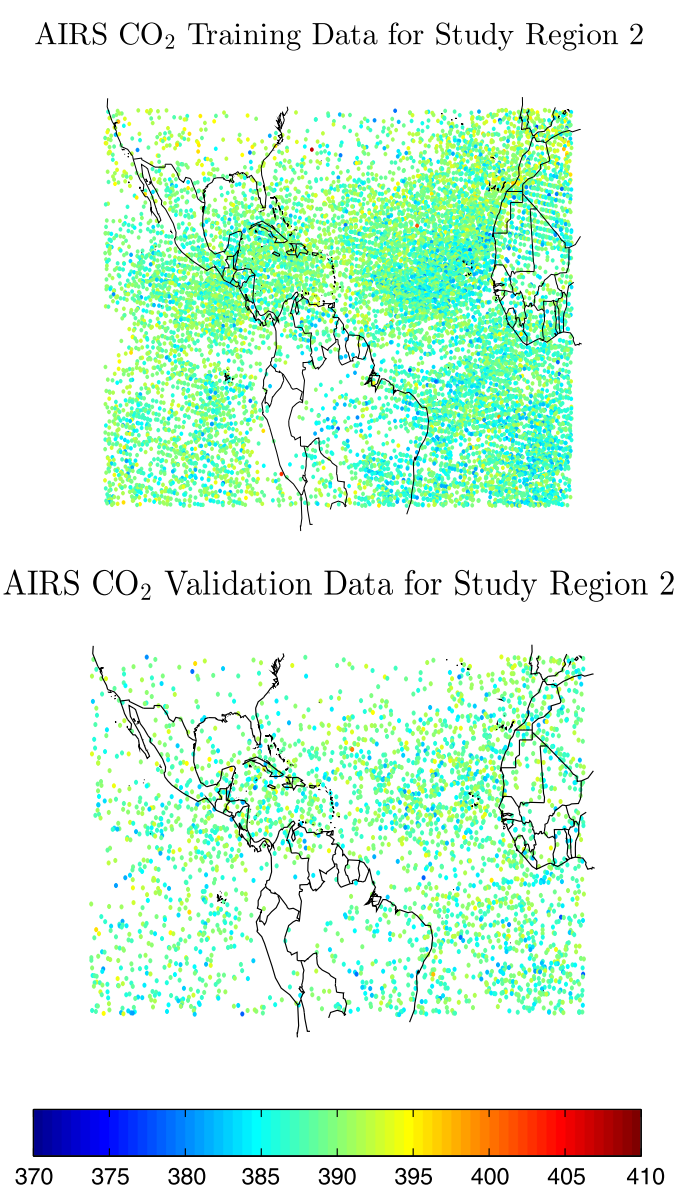

FIG 2. *A spatial dataset made up of 9 days of measurements of mid-tropospheric* CO_2 *in parts per million (ppm). The data considered are between* $-20°$ *degrees and* $44°$ *degrees latitude and* $-125°$ *degrees and* $3°$ *degrees longitude, from Februrary 1 through Februrary 9, 2010. The data are randomly split into training and validation datasets with* $n = 12,358$ *and* $m = 3,090$*, respectively.*

and latitudes $-20°$ to $44°$ (this is the same study region used in 25). There are $n = 12,358$ observations used to train each spatial predictor (top panel of Figure 2), and $m = 3,090$ observations used for validation (bottom panel of

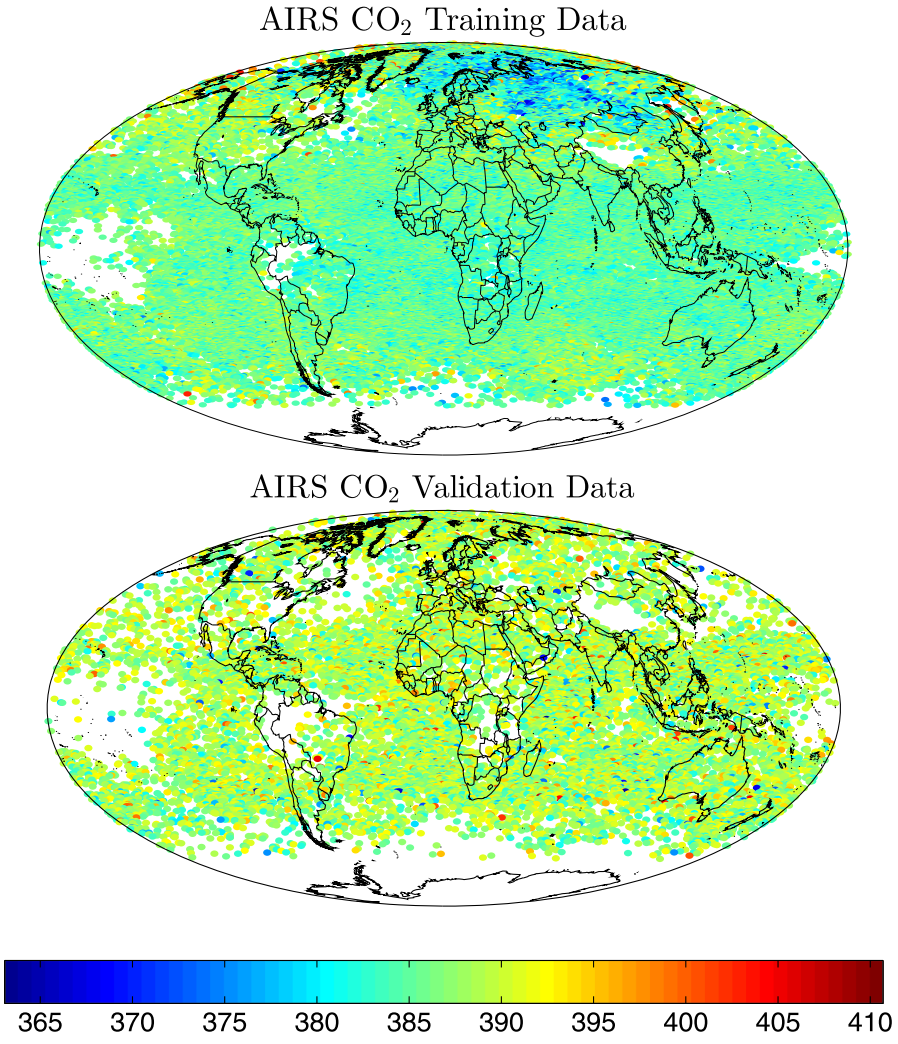

FIG 3. *A spatial dataset made up of 9 days of measurements of global mid-tropospheric* CO_2 *in parts per million (ppm). The data considered are between* $-60°$ *degrees and* $90°$ *degrees latitude from Februrary 1 through Februrary 9, 2010. The data are randomly split into training and validation datasets with* $n = 44,621$ *and* $m = 2,000$*, respectively.*

Figure 2). In Section 4.2, we use the data in Figure 2 to compare these five spatial predictors.

Finally, in Section 4.3, we use the entire dataset in Figure 3, which is computationally feasible only for EDW, FRK, SPD, and LTK, but no longer for MPP. There are $n = 59,488$ observations used to train each spatial predictor (top panel of Figure 3), and $m = 14,873$ observations used for validation (bottom panel of Figure 3). This is by no means an unusually large dataset that one might process spatially; for example, Sengupta et al. [41] and Bradley et al. [5]

process datasets on the order of millions. Note that in Appendix B, we consider different specifications of Study Region 3. This is done in an effort to determine the effect that the training and validation datasets have on our conclusions.

The training (validation) data are referenced by their locations, $D^{\text{trn}} \equiv \{\mathbf{s}_j^{\text{trn}} : j = 1, \ldots, n\}$ ($D^{\text{val}} \equiv \{\mathbf{s}_j^{\text{val}} : j = 1, \ldots, m\}$), where $D_O = D^{\text{trn}} \cup D^{\text{val}}$ and $D^{\text{trn}} \cap D^{\text{val}} = \emptyset$. Hence, the total size of the dataset is $n + m$. We use the validation datasets to assess the predictive performance of each spatial predictor. Define the root average squared testing error (RSTE) associated with the predictor $\hat{Y}^{\text{PRD}}$ and the different study regions as,

$$\text{RSTE}(\hat{Y}^{\text{PRD}}, m) \equiv \left(\frac{1}{m} \sum\nolimits_{j=1}^{m} (Z(\mathbf{s}_j^{\text{val}}) - \hat{Y}^{\text{PRD}}(\mathbf{s}_j^{\text{val}}, \mathbf{Z}))^2 \right)^{1/2}, \tag{4.1}$$

where "PRD" notates a generic predictor. The RSTE will be used to compare each of the seven spatial predictors (small values are desirable), PRD = TSK, SSP, EDW, FRK, MPP, SPD, and LTK.

Another criterion that we consider is the predictive model choice criterion (PMCC) from Gneiting and Raftery [20, see their Equation (27)],

$$\text{PMCC}(\hat{Y}^{\text{PRD}}, m) \equiv \frac{1}{m} \sum\nolimits_{j=1}^{m} \frac{(Z(\mathbf{s}_j^{\text{val}}) - \hat{Y}^{\text{PRD}}(\mathbf{s}_j^{\text{val}}, \mathbf{Z}))^2}{\hat{\sigma}^{\text{PRD}}(\mathbf{s}_j^{\text{val}}, \mathbf{Z})^2} - \log\left(\hat{\sigma}^{\text{PRD}}(\mathbf{s}_j^{\text{val}}, \mathbf{Z})^2\right), \tag{4.2}$$

where $\hat{\sigma}^{\text{PRD}}(\,\cdot\,,\,\cdot\,)^2$ is the model-based posterior variance, and hence we can only compute the PMCC for PRD = TSK, FRK, MPP, SPD, and LTK. Notice that for the SME model in (3.1) and PRD = TSK, FRK, and LTK,

$$\begin{aligned}\hat{\sigma}^{\text{PRD}}(\mathbf{s}, \mathbf{Z})^2 = {} & \text{var}(\nu(\mathbf{s})|\hat{\boldsymbol{\theta}}^{\text{PRD}}) + \text{var}(\xi(\mathbf{s})|\hat{\boldsymbol{\theta}}^{\text{PRD}}) \\ & - \text{cov}(\mathbf{Z}, Y(\mathbf{s})|\hat{\boldsymbol{\theta}}^{\text{PRD}})' \text{cov}(\mathbf{Z}|\hat{\boldsymbol{\theta}}^{\text{PRD}})^{-1} \text{cov}(\mathbf{Z}, Y(\mathbf{s})|\hat{\boldsymbol{\theta}}^{\text{PRD}}).\end{aligned}$$

The posterior variance for the predictors that are derived using a fully Bayesian approach are estimated by

$$\hat{\sigma}^{\text{PRD}}(\mathbf{s}, \mathbf{Z})^2 = \begin{cases} \text{var}\left(Z(\mathbf{s})_\ell : \ell = 1, \ldots, L\right) & \text{if PRD} = \text{MPP}, \\ \text{var}\left(Y(\mathbf{s})_\ell : \ell = 1, \ldots, L\right) & \text{if PRD} = \text{SPD}; \;\; \mathbf{s} \in D^{\text{val}}, \end{cases}$$

where recall $\{Z(\cdot)_\ell\}$ and $\{Y(\cdot)_\ell\}$ are samples from their respective posterior distributions defined in Sections 3.5 and 3.6, respectively. The PMCC is useful for comparing predictors (small values are desirable) because it incorporates information on the implicit model-based prediction errors. However, it has the limitation of not allowing a comparison to deterministic predictors, something that can be done with RSTE.

We are interested in evaluating other properties of the predictors in addition to their predictive performance. In particular, to assess the amount of smoothness in PRD, consider the lag-1 semivariogram,

$$\frac{1}{2|C(1)|} \sum_{C(1)} (\hat{Y}^{\text{PRD}}(\mathbf{u}_i) - \hat{Y}^{\text{PRD}}(\mathbf{u}_j))^2, \tag{4.3}$$

where PRD = TSK, SSP, EDW, FRK, MPP, SPD, and LTK, $C(h) \equiv \{(i,j) : ||\mathbf{u}_i - \mathbf{u}_j|| = h\}$, $|C(h)|$ denotes the number of distinct elements in the set $C(h)$, h denotes the spatial lag, and $h = 1$ is in a unit of distance defined by the smallest lag at which a semivariogram can be computed. In Study Regions 1, 2, and 3, the unit of distance is 1.41°, 1.5°, and 1.5°, respectively. A large (small) lag-1 semivariogram in (4.3) suggests that the map of PRD is non-smooth (smooth).

The exact specifications of each of the seven spatial predictors can be found in Section 3. Here the covariates are $\mathbf{x}((\text{latitude}, \text{longitude})') \equiv (1, \text{latitude})$, since it is well known that mid-tropospheric CO_2 values display a latitudinal gradient [21]; that is, there are $p = 2$ covariates. Additionally, the measurement-error variances are assumed known for TSK, FRK, and LTK; in practice, these variances are estimated using a variogram-extrapolation technique used by Kang et al. [25] and Katzfuss and Cressie [28]. We use their estimate of $\sigma^2_\epsilon = 5.6062$ ppm^2 and, for simplicity, we shall take $V_\epsilon(\cdot) \equiv 1$.

In Sections 4.1 through 4.3, all of our computations are performed on a Dell Optiplex 7010 Desktop Computer with a quad-Core 3.40 GHz processor and 8 Gbytes of memory. It is important to note that the timing and memory-usage results may be different for different machines; however, to illustrate what someone might expect in practice, we use a computer that has the specification of a "typical personal desktop." Additionally, FRK, EDW, and SSP were computed using Matlab, which has built-in multithreading capabilities that allows all four cores to be used; however, TSK, MPP, SPD, and LTK are written in R, and thus they use a single core.

4.1. Comparison using a small dataset of mid-tropospheric $\mathbf{CO_2}$

In this section, we use the data in Study Region 1 (see Figure 1), which we process using all seven spatial predictors, namely $\hat{Y}^{\text{TSK}}$, $\hat{Y}^{\text{SSP}}$, $\hat{Y}^{\text{EDW}}$, $\hat{Y}^{\text{FRK}}$, $\hat{Y}^{\text{MPP}}$, $\hat{Y}^{\text{SPD}}$, and $\hat{Y}^{\text{LTK}}$. Maps of the seven spatial predictors for Study Region 1 are given in Figure 4.

Each spatial predictor displays similar general patterns, with lower CO_2 values near the Great Lakes. In general, we can separate the predictors in Figure 4 into two categories: smooth and non-smooth. The two deterministic predictors (SSP and EDW) appear non-smooth, whereas the stochastic spatial predictors appear quite smooth; this is also seen in the lag-1 semivariograms in Table 1. This may be because the stochastic predictors rely on an underlying smooth covariance function in this setting, where the dataset is small and fairly sparse over the study region.

The RSTE results for this example (given in Table 1) indicate that FRK is the predictor that has the most-favorable predictive performance, while LTK has the least-favorable predictive performance among the seven spatial predictors; however, it should be noted that the RSTE values are fairly similar across different choices of PRD. The PMCC results for this example (given in Table 1) indicate that TSK and MPP have the most-favorable predictive performance, followed by FRK, and SPD and LTK have the least-favorable predictive performance among the five stochastic spatial predictors; recall that PMCC is not

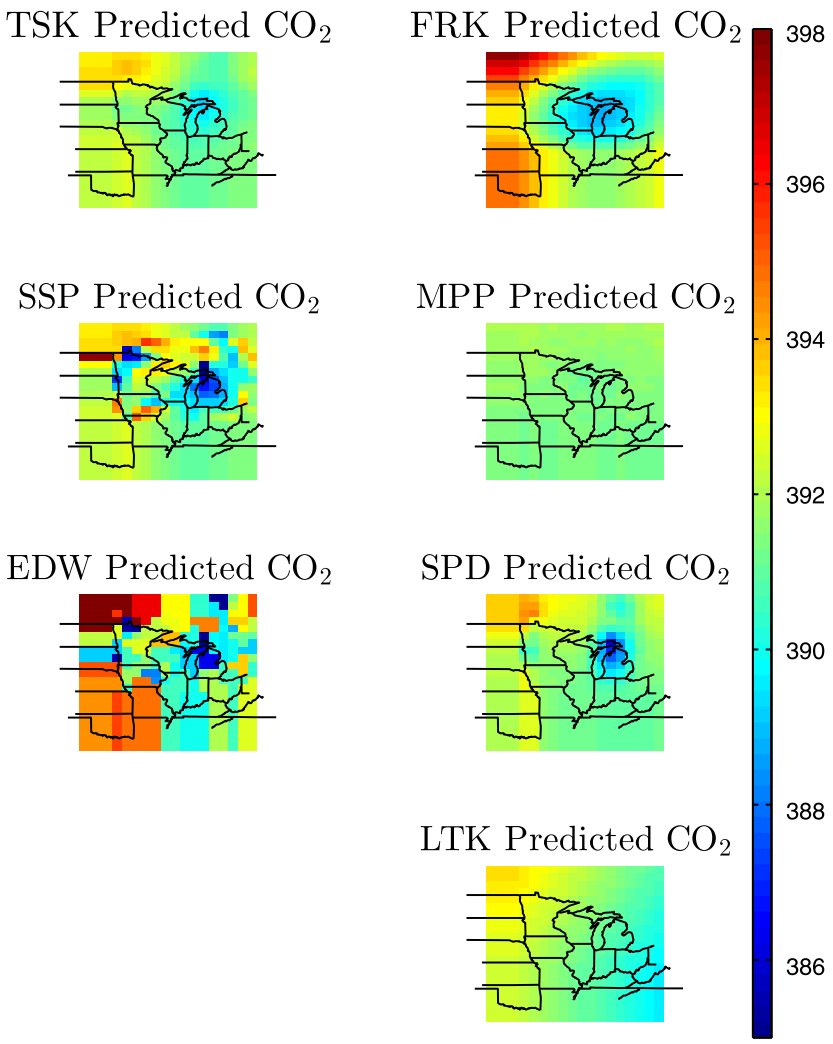

FIG 4. *Spatial prediction (in ppm) of mid-tropospheric* CO_2 *concentrations using TSK, SSP, EDW, FRK, MPP, SPD, and LTK. Predictions are indicated in the title headings and are mapped over Study Region 1.*

TABLE 1

Results from Study Region 1 (Section 4.1) for the root average squared testing error (RSTE), PMCC, lag-1 semivariogram, CPU time (including model fitting and parameter estimation), and peak memory-usage by predictor. These quantities are produced using the data shown in Figure 1.

Predictor	RSTE	PMCC	Lag-1 Semivariogram	CPU Time (in minutes)	Peak Memory-Usage (in MB)
TSK	4.7063	-0.4845	0.5739	0.20	171.08
SSP	4.7151	N/A	4.9746	0.02	1,043.80
EDW	4.5126	N/A	7.8176	2.31	889.54
FRK	4.3097	12.5612	2.1298	1.01	791.12
MPP	4.9084	-0.5873	0.0339	3.37	239.51
SPD	4.7399	26.2548	1.1271	0.24	143.14
LTK	5.0163	39.5806	0.2536	2.73	205.84

defined for deterministic predictors. As expected, there were no difficulties with CPU time and memory usage for this small dataset, and each of the seven spatial predictors were computed in a matter of seconds.

4.2. Comparison using a large dataset of mid-tropospheric $\mathbf{CO_2}$

It is well known that the inversion of a large $n \times n$ matrix makes TSK and SSP computationally impractical. Hence, for this large dataset in Study Region 2 (see Figure 2) we consider the five spatial predictors that can be computed, namely EDW, FRK, MPP, LTK, and SPD.

Maps of the five spatial predictors $\hat{Y}^{\text{EDW}}$, $\hat{Y}^{\text{FRK}}$, $\hat{Y}^{\text{MPP}}$, $\hat{Y}^{\text{SPD}}$, and $\hat{Y}^{\text{LTK}}$ are given in Figure 5. Each spatial predictor displays similar general patterns; in contrast to the results in Section 4.1, the large dataset used in this section shows clearly that MPP is the smoothest predictor, EDW is the least smooth, and FRK, LTK, and SPD have similar patterns of smoothness. These results are further corroborated by inspecting the lag-1 semivariograms in Table 2.

The RSTE results for this example (see Table 2) are fairly constant across different choices of PRD, with MPP (EDW) having the most-favorable (least-favorable) predictive performance as measured by RSTE; recall that MPP is the smoothest spatial predictor. Similar conclusions can be made from the PMCC results in Table 2, where the reduced-rank prediction methods (i.e., FRK and MPP) appear to have the most-favorable predictive performances, and the full-rank prediction methods (i.e., EDW, LTK, and SPD) appear to have less-favorable predictive performances. Notice that for the small dataset (in Section 4.1), some computation times are slightly larger than those for the large dataset. This is caused by the iterative estimation of parameters in FRK and LTK in small-data situations. The CPU time and memory usage are manageable except for MPP, which has a CPU time of approximately 3.5 hours.

4.3. Comparison using a very large dataset of mid-tropospheric $\mathbf{CO_2}$

In this section, we use the data in Study Region 3 (the entire dataset; see Figure 3), and the four spatial predictors that can process a dataset of this size; that is, we compare EDW, FRK, SPD, and LTK. Note that the MPP predictor, which is computed using a Metropolis-within-Gibbs sampler, is too computationally intensive for very large spatial datasets and already had difficulty with the large dataset in Section 4.2. Notice that the four spatial predictors that can handle datasets of this size do not use MCMC algorithms for statistical inference; specifically, FRK and LTK are empirical Bayesian, SPD uses a fully Bayesian approach based on Rue et al. [38]'s INLA algorithm, and EDW does not use a statistical model at all for inference.

Maps of the four spatial predictors, $\hat{Y}^{\text{EDW}}$, $\hat{Y}^{\text{FRK}}$, $\hat{Y}^{\text{SPD}}$, and $\hat{Y}^{\text{LTK}}$ are given in Figure 6. As was the case for Sections 4.1 and 4.2, each prediction method displays similar general patterns. The lag-1 semivariograms indicate that SPD

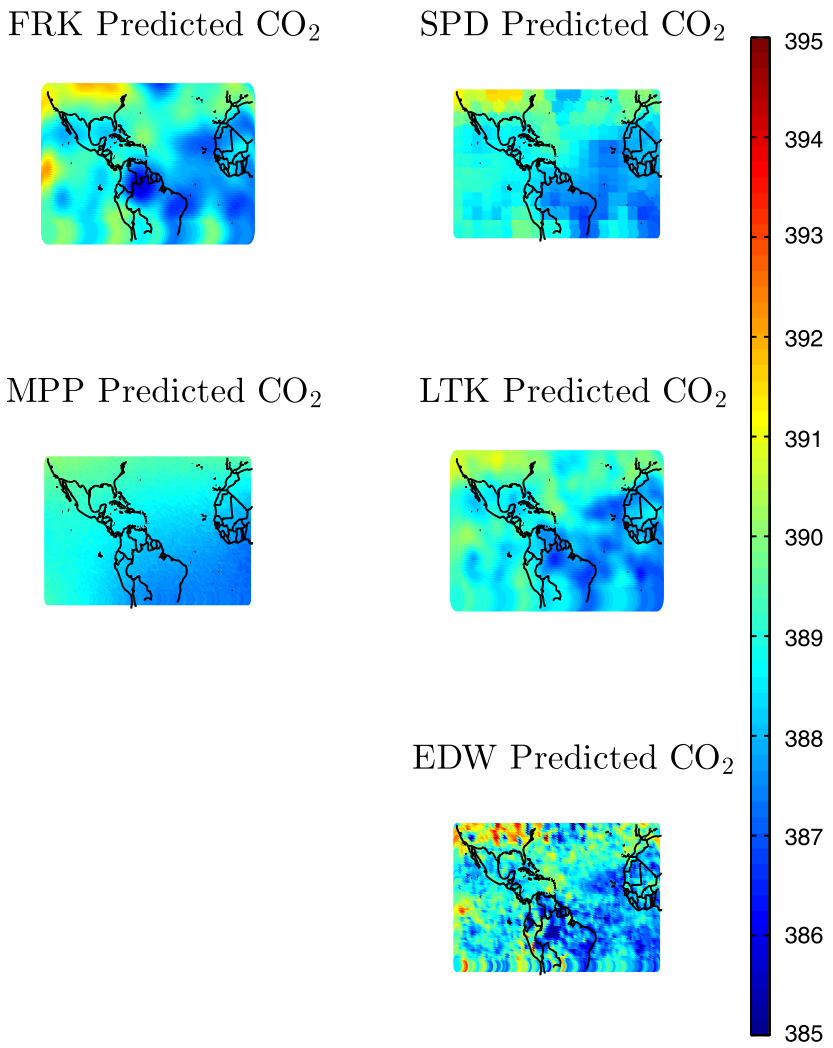

FIG 5. *Spatial prediction (in ppm) of mid-tropospheric* CO_2 *concentrations using FRK, MPP, SPD, LTK, and EDW. Predictions are indicated in the title headings and are mapped over Study Region 2.*

TABLE 2

Results from Study Region 2 (Section 4.2) for the root average squared testing error (RSTE), PMCC, lag-1 semivariogram, CPU time (including model fitting and parameter estimation), and peak memory-usage by predictor. These quantities are produced using the data shown in Figure 2.

Predictor	RSTE	PMCC	Lag-1 Semivariogram	CPU Time (in minutes)	Peak Memory-Usage (in MB)
EDW	3.0811	N/A	0.8167	28.49	910.4589
FRK	3.0067	12.6155	0.2075	0.52	841.0030
MPP	2.9243	-1.0327	0.0164	216.79	2042.6
SPD	2.9630	70.0529	0.1243	0.47	111.18
LTK	2.9855	27.3636	0.1470	1.72	1,971.8

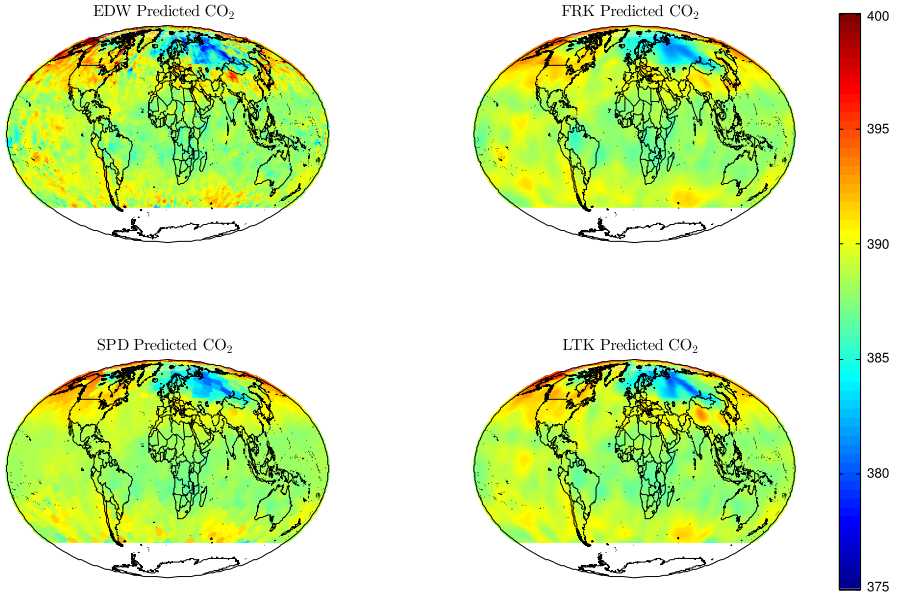

FIG 6. *Spatial prediction (in ppm) of mid-tropospheric* CO_2 *concentrations using EDW, SPD, FRK, and LTK. Predictions are indicated in the title headings and are mapped over Study Region 3. Note that there is no prediction given below latitude* $-60°$*, since AIRS has not released any observations there.*

TABLE 3

Results from Study Region 3 (Section 4.3) for the root average squared testing error (RSTE), PMCC, lag-1 semivariogram, CPU time (including model fitting and parameter estimation), and peak memory-usage by predictor. These quantities are produced using the data shown in Figure 3.

Predictor	RSTE	PMCC	Lag-1 Semivariogram	CPU Time (in minutes)	Peak Memory-Usage (in MB)
EDW	5.5203	N/A	1.5978	279.74	939.42
FRK	3.9841	12.0974	0.5080	0.51	1,025.40
SPD	3.9882	53.1760	2.1121	4.72	165.19
LTK	4.0026	45.1762	0.1440	85.13	490.60

is now the least smooth among the four predictors; LTK retains its property of being much smoother than FRK, EDW, and SPD.

The RSTE results for this example (see Table 3) are fairly constant across different choices of PRD (similar to the results in Sections 4.1 and 4.2), with FRK (EDW) having the most-favorable (least-favorable) predictive performance as measured by RSTE. As in Sections 4.1 and 4.2, the PMCC results show that a reduced-rank method, FRK, has more-favorable predictive performance than the full-rank methods, SPD and LTK. The CPU time for both FRK and SPD indicate that both of these methods are highly computationally efficient for spatial prediction. Moreover, the memory usage for each predictor is modest. However, EDW and LTK required substantial CPU time to obtain spatial predictions (around 1.5 hours and 4.7 hours, respectively).

5. Discussion

In this article, we present a comparison of spatial predictors from an algorithmic viewpoint. In particular, we systematically lay out the parameterization, the algorithm, and the motivation of three traditional methods of spatial prediction and four more-recently-introduced spatial predictors. The traditional spatial predictors include: traditional stationary kriging (TSK), smoothing splines (SSP), and negative-exponential distance-weighting (EDW). The more-recently-introduced spatial predictors include: fixed rank kriging (FRK), a modified predictive processes approach (MPP), a stochastic partial differential equation approach (SPD), and lattice kriging (LTK). Additionally, we use a benchmark of small, large, and very large mid-tropospheric CO_2 datasets to compare computation time, memory-usage, and the prediction performance of each spatial predictor. For researchers who have other spatial predictors to add to our list of seven, we have made training and validation datasets available in the Supplemental Material, making it easy for new comparisons to be made.

Recent advances in technology, such as remote sensing, have made large-to-massive spatial datasets more available, making spatial prediction with "big spatial data" an important and growing problem in the statistics literature. Consequently, the algorithmic concerns of CPU time and memory-usage are featured in our comparison along with predictive performance.

Of the seven predictors we consider, FRK and SPD perform extremely well in terms of CPU time and memory-usage. However, the remaining five spatial predictors are not as efficient. Both EDW and LTK can be used for the very large benchmark dataset, but the CPU time was quite long (approximately 1.5 and 4.7 hours, respectively). It is well known that TSK and SSP have very poor CPU time and memory-usage properties for large datasets and, hence, we were only able to use these predictors on the small benchmark dataset. The MPP predictor also has limitations in CPU time; consequently, we were only able to use MPP on the small and large benchmark datasets, the latter dataset resulting in a significant CPU time (around 3.5 hours).

When visually comparing each of the seven spatial predictors, we see that they each display similar general patterns. From an algorithmic point-of-view, this is to be expected, since if the signal-to-noise ratio is "large enough," then any local-averaging scheme should be able to find the large-scale patterns. These visual patterns are further corroborated using the lag-1 semivariogram, which is consistently smaller (larger) for MPP (EDW and SPD). Of the three stochastic predictors that can be computed for all study regions, FRK has more favorable predictive performance than SPD and LTK, according to the PMCC criterion. Of the four predictors that can be computed for all study regions, EDW had the least-favorable predictive performance (among FRK, SPD, LTK, and EDW), according to the RSTE criterion.

It is important to point out that each spatial predictor is motivated differently and, consequently, there are clear differences in terms of the interpretation of parameters. Our exposition has focused on a comparison of predictors rather than on a comparison of models and parameters. However, it would also be

interesting to compare estimates of parameters that have a unique meaning across different predictors. For example, MPP, LTK, and TSK each have a (equivalent) range parameter that could be estimated and compared.

Ultimately, which spatial predictor is "best" depends heavily on the dataset. We suggest that the practical comparisons given in Section 4 be used to guide which subset of predictors to use. In the case of the benchmark AIRS CO_2 dataset, empirical comparisons between reduced-rank and full-rank spatial predictors in Section 4 shed new light on the recent criticisms of reduced-rank statistical modeling [30, 43]. Reduced-rank methods have done well in a number of settings [see, e.g., 2, 5, 12, 13, 14, 17, 24, 25, 27, 28, 34, 42, 48]. In terms of predictive performance as measured by RSTE and PMCC, our results on a benchmark dataset of CO_2 data from NASA's AIRS instrument showed that reduced-rank methods outperform the viable full-rank alternatives.

The training and validation datasets for the very large mid-tropospheric CO2 dataset are available in the Supplemental Material.

Appendix A: Matlab code for negative-exponential distance-weighting

Algorithm 1: Matlab code to compute EDW

```
% % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % %
% Summary: The Matlab function used to compute the EDW predictor from Section 3.3.
% % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % %
% Inputs:
% train_data: The n-dimensional vector of training data.
% train_locs: The n x 2 matrix of the training data's lat/lon coordinates.
% thetaEDW: The positive real-valued smoothing parameter defined in Section 3.3.
% pred_locs: The m x 2 matrix consisting of the lat/lon coordinates associated with
% the prediction locations.
% % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % %
% Outputs:
% YhatEDW: The m-dimensional EDW predictor.
% % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % %
function [YhatEDW] = EDW(train_data,train_locs,thetaEDW,pred_locs)
m = size(pred_locs,1);
n = size(train_data,1);
YhatEDW = zeros(m,1);
    for i = 1:m
        total = 0;
            for j = 1:n
                temp1 = sqrt(sum((pred_locs(i,:) - train_locs(j,:)).^2));
                temp = exp(-thetaEDW*temp1);
                total =  temp+total;
                YhatEDW(i) = train_data(j)*temp + YhatEDW(i);
            end
    disp(i);
    YhatEDW(i) = YhatEDW(i)/total;
    end
end
```

TABLE 4
Entries in the table show root average squared testing error (RSTE) results based on the very large dataset in Figure 3 (referred to as TVD 1), and based on two other random selections of training and validation data (referred to as TVD 2 and TVD 3, respectively).

Predictor	TVD 1	TVD 2	TVD 3
EDW	5.5203	5.7155	5.4825
FRK	3.9841	3.9891	3.9546
SPD	3.9882	3.9768	3.9898
LTK	4.0026	3.9923	3.9979

Appendix B: Hold-out sensitivity analysis

The relative differences of RSTE in Tables 1–3 are very small, indicating that the predictive performance of each of these predictors are comparable. This is notable considering a recent criticism of reduced-rank approaches [43]. These small differences in RSTE motivated us to determine whether or not the differences in RSTE can be explained by the randomization used to divide up the total data available into a training dataset and a validation dataset.

Thus, we considered different randomly generated training and validation datasets for Study Region 3. We refer to the training and validation data in Figure 3 as TVD 1, where "TVD" stands for "training and validation datasets." Then, we considered two other random generations of training and validation datasets, which are referred to as TVD 2 and TVD 3, respectively. In Table 4, we give the RSTE values by TVD. In general, we see that the relative rankings of FRK to SPD may be due to the randomization of data into training and validation; however, we consistently see that LTK performs worse than FRK and SPD, and that EDW performs worse than all three of its competitors.

References

[1] BANERJEE, S., CARLIN, B. P., and GELFAND, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. London, UK: Chapman and Hall.

[2] BANERJEE, S., GELFAND, A. E., FINLEY, A. O., and SANG, H. (2008). "Gaussian predictive process models for large spatial data sets." *Journal of the Royal Statistical Society Series B*, 70, 825–848. MR2523906

[3] BRADLEY, J. R., CRESSIE, N., and SHI, T. (2011). "Selection of rank and basis functions in the Spatial Random Effects model." In *Proceedings of the 2011 Joint Statistical Meetings*, 3393–3406. Alexandria, VA: American Statistical Association.

[4] BRADLEY, J. R., CRESSIE, N., and SHI, T. (2015). "Comparing and selecting spatial predictors using local criteria (with discussion)." *TEST*, 24, 1–28 (Rejoinder: pp. 54 – 60). MR3314567

[5] BRADLEY, J. R., HOLAN, S. H., and WIKLE, C. K. (2015). "Multivariate spatio- temporal models for high-dimensional areal data with application to Longitudinal Employer-Household Dynamics." *The Annals of Applied Statistics*, 9, 1761–1791. MR3456353

[6] Campbell, J. B. (2010). "Improving lead generation success through integrated methods: transcending 'drug discovery by numbers'." *IDrugs: the Investigational Drugs Journal*, 21, 62–71.

[7] Chahine, M. T., Pagano, T. S., Aumann, H. H., Atlas, R., Barnet, C., Blaisdell, J., Chen, L., Divakarla, M., Fetzer, E. J., Goldberg, M., Gautier, C., Granger, S., Hannon, S., Irion, F. W., Kakar, R., Kalnay, E., Lambrigtsen, B. H., Lee, S. Y., Marshall, J. L., McMillian, W. W., McMillin, L., Olsen, E. T., Revercomb, H., Rosenkranz, P., Smith, W. L., Staelin, D., Strow, L. L., Susskind, J., Tobin, D., Wolf, W., and Zhou, L. (2006). "AIRS: Improving weather forecasting and providing new data on greenhouse gases." *Bulletin of the American Meteorological Society*, 87, 911–926.

[8] Chiles, J. P. and Delfiner, P. (1999). *Geostatistics: Modeling Spatial Uncertainty*. Hoboken, NJ: Wiley. MR1679557

[9] Cressie, N. (1988). "Spatial prediction and ordinary kriging." *Mathematical Geology*, 20, 405–421. MR0950457

[10] Cressie, N. (1990). "The origins of kriging." *Mathematical Geology*, 22, 239–252. MR1047810

[11] Cressie, N. (1993). *Statistics for Spatial Data,* rev. edn. New York, NY: Wiley. MR1239641

[12] Cressie, N. and Johannesson, G. (2006). "Spatial prediction for massive data sets." In *Australian Academy of Science Elizabeth and Frederick White Conference*, 1–11. Canberra: Australian Academy of Science.

[13] Cressie, N. and Johannesson, G. (2008). "Fixed rank kriging for very large spatial data sets." *Journal of the Royal Statistical Society, Series B*, 70, 209–226. MR2412639

[14] Cressie, N., Shi, T., and Kang, E. L. (2010). " Fixed Rank Filtering for spatio-temporal data." *Journal of Computational and Graphical Statistics*, 19, 724–745. MR2732500

[15] Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. Hoboken, NJ: Wiley. MR2848400

[16] Finley, A. O. and Banerjee, S. (2013). *spBayes: Univariate and Multivariate Spatial-temporal Modeling*. R package version 0.3-8.

[17] Finley, A. O., Sang, H., Banerjee, S., and Gelfand, A. E. (2009). "Improving the performance of predictive process modeling for large datasets." *Computational Statistics and Data Analysis*, 53, 2873–2884. MR2667597

[18] Gandin, L. S. (1965). *Objective Analysis of Meteorological Fields: Gidrometeorologicheskoe Izdatel'stvo (GIMIZ), Leningrad*. Jerusalem: translated by Israel Program for Scientific Translations.

[19] Gauss, C. (1809). *Theoria motus corporum celestium: Perthes et Besser, Hamburg*. Translated as "Theory of motion of the heavenly bodies moving about the sun in conic sections", trans. C. H. Davis. Boston, MA: Little, Brown.

[20] Gneiting, T. and Raftery, A. (2007). "Strictly proper scoring rules, pre-

diction, and estimation." *Journal of the American Statistical Association*, 102, 359–378. MR2345548

[21] Hammerling, D. M., Michalak, A. M., and Kawa, S. R. (2012). "Mapping of CO2 at high spatiotemporal resolution using satellite observations: Global distributions from OCO-2." *Journal of Geophysical Research*, 117, 1–10.

[22] Henderson, H. V. and Searle, S. R. (1981). "On deriving the inverse of a sum of matrices." *SIAM Review*, 23, 53–60. MR0605440

[23] Hormozi, A. M. and Giles, S. (2004). "Data mining: A competitive weapon for banking and retail industries." *Information Systems Management*, 21, 62–71.

[24] Kang, E. L. and Cressie, N. (2011). "Bayesian inference for the Spatial Random Effects model." *Journal of the American Statistical Association*, 106, 972–983. MR2894757

[25] Kang, E. L., Cressie, N., and Shi, T. (2010). "Using temporal variability to improve spatial mapping with application to satellite data." *Canadian Journal of Statistics*, 38, 271–289. MR2682762

[26] Katzfuss, M. and Cressie, N. (2009). "Maximum likelihood estimation of covariance parameters in the spatial-random-effects model." In *Proceedings of the Joint Statistical Meetings*, 3378–3390. Alexandria, VA: American Statistical Association.

[27] Katzfuss, M. and Cressie, N. (2011). "Spatio-temporal smoothing and EM estimation for massive remote-sensing data sets." *Journal of Time Series Analysis*, 32, 430–446. MR2841794

[28] Katzfuss, M. and Cressie, N. (2012). "Bayesian hierarchical spatio-temporal smoothing for very large datasets." *Environmetrics*, 23, 94–107. MR2873787

[29] Li, X., Cheng, G., and Lu, L. (2000). "Comparison of spatial interpolation methods." *Advances in Earth Sciences*, 15, 260–265.

[30] Lindgren, F., Rue, H., and Lindström, J. (2011). "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach." *Journal of the Royal Statistical Society, Series B*, 73, 423–498. MR2853727

[31] Matérn, B. (1960). "Spatial Variation." *Meddelanden fran Statens Skogsforskningsinstitut*, 49, 1–144.

[32] Matheron, G. (1963). "Principles of geostatistics." *Economic Geology*, 58, 1246–1266.

[33] Nguyen, H., Cressie, N., and Braverman, A. (2012). "Spatial statistical data fusion for remote sensing applications." *Journal of the American Statistical Association*, 107, 1004–1018. MR3010886

[34] Nguyen, H., Katzfuss, M., Cressie, N., and Braverman, A. (2014). "Spatio-temporal data fusion for remote-sensing applications." *Technometrics*, 56, 174–185. MR3207845

[35] Nychka, D. (2001). "Spatial process estimates as smoothers." In *Smoothing and Regression: Approaches, Computation and Applications,* rev. edn, ed. M. G. Schmiek, 393–424. New York, NY: Wiley.

[36] Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., and Sain, S. (2015). "A multi-resolution Gaussian process model for the analysis of large spatial data sets." *Journal of Computational and Graphical Statistics*, 2, 579–599. MR3357396

[37] Ribeiro, P. J. and Diggle, P. J. (2001). "geoR: a package for geostatistical analysis." *R-NEWS*, 1, 2, 14–18. ISSN 1609-3631.

[38] Rue, H., Martino, S., and Chopin, N. (2009). "Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations." *Journal of the Royal Statistical Society, Series B*, 71, 319–392. MR2649602

[39] Rue, H., Martino, S., Lindgren, F., Simpson, D., and Riebler, A. (2014). *INLA: Functions which allow to perform full Bayesian analysis of latent Gaussian models using Integrated Nested Laplace Approximaxion*. R package version 3.0.2.

[40] Schabenberger, O. and Gotway, C. (2005). *Statistical Methods for Spatial Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC Press. MR2134116

[41] Sengupta, A., Cressie, N., Frey, R., and Kahn, B. (2012). "Statistical modeling of MODIS cloud data using the Spatial Random Effects model." In *Proceedings of the Joint Statistical Meetings*, 3111–3123. Alexandria, VA: American Statistical Association.

[42] Shi, T. and Cressie, N. (2007). "Global statistical analysis of MISR aerosol data: A massive data product from NASA's Terra satellite." *Environmetrics*, 18, 665–680. MR2408937

[43] Stein, M. (2014). "Limitations on low rank approximations for covariance matrices of spatial data." *Spatial Statistics*, 8, 1–19. MR3326818

[44] Sun, Y., Li, B., and Genton, M. G. (2012). "Geostatistics for large datasets." In *Space-Time Processes and Challenges Related to Environmental Problems*, eds. E. Porcu, J. M. Montero, and M. Schlather, 55–77. Berlin, DE: Springer.

[45] Tierney, L. and Kadane, J. B. (1986). "Accurate approximations for posterior moments and marginal densities." *Journal of the American Statistical Association*, 81, 82–86. MR0830567

[46] Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia, PA: Society for Industrial and Applied Mathematics. MR1045442

[47] Whittle, P. (1963). "Stochastic processes in several dimensions." *Bulletin of the International Statistical Institute*, 40, 974–994. MR0173287

[48] Wikle, C. K. and Cressie, N. (1999). "A dimension-reduced approach to space-time Kalman filtering." *Biometrika*, 86, 815–829. MR1741979

[49] Wikle, C. K., Milliff, R. F., Nychka, D., and Berliner, L. M. (2001). "Spatiotemporal hierarchical Bayesian modeling of tropical ocean surface winds." *Journal of the American Statistical Association*, 96, 382–397. MR1939342